Network Project
A Growing Network Model

CID: 01874240

27th March 2023

Abstract: In this project, the Barabási-Albert network model governed by the original Preferential Attachment (PA), by Random Attachment (RA), and by Existing Vertices (EV) rules is investigated. The network is simulated in Python using Object-Oriented Programming (OOP). For all models, theoretical degree probability distributions in the long time limit are derived. To check the quality of the theory, statistical tests were performed. In the PA and RA models, the numerical results agree with theory, as the p value for regions before the characteristic bump is 1.000 for any chosen parameters. The characteristic bumps are present in all models due to the finite-size effects. The EV model showed the worst agreement with the theory as the p value for larger m was less then 0.05, even in the regions before the bump. A theoretical scaling of the largest degree was also derived in all models. This was used to perform a data collapse by aligning the bumps. Out of three, the PA model showed the best results, while the EV model was the worst. Due to the intricacies of the EV model, a better theoretical foundation is needed.

Word Count: 2489

0 Introduction

The Barabási-Albert network model (BA) is a mathematical model which describes the relational data. One of its properties is the scale-free behaviour of such networks: the vertices connectivities can be described by a power-law. The scale-free behaviour also indicates that large networks exhibit self-organising phenomena. This leads to a potential data collapse, which this project aims to achieve.

In this project, the BA model is simulated in three different ways: the network being governed by the original Preferential Attachment (PA), by Random Attachment (RA), and by Existing Vertices (EV) rules.

Definition

In this project, the BA network evolves in time by adding a new node and connecting m edges to it. The other ends of these edges are connected to existing nodes with a probability determined by the model type. A node's degree is defined as the number of edges connected to it.

In the PA model, the new node is connected to m existing nodes with a probability proportional to their degree, k:

$$\Pi_{PA}(k,t) = \frac{k}{2E(t)}, \tag{0.0.1}$$

where $E(t) = \frac{1}{2}\sum_k k$ is the total number of edges.

In the RA model, the probability of attaching a new node to an existing vertex is equal for all vertices:

$$\Pi_{RA}(k,t) = \frac{1}{N(t)}, \tag{0.0.2}$$

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

In the EV model, the new edges are attached in the following way:

- r edges are attached between a new node and the existing nodes with probability Π_{RA}.
- $m - r$ edges are attached between the existing nodes with probability Π_{PA}.

For the purposes of this project only $r = m/3$ cases are investigated.

1 Phase 1: Pure Preferential Attachment Π_{pa}

1.1 Implementation

1.1.1 Numerical Implementation

To implement the PA probability, a list of edges' ends is created. It automatically contains information about nodes' degrees. `random` chooses an end with random probability, but chooses a node with preferential probability. While choosing the nodes, a temporary list of new ends is created to ensure that only one edge is allowed between 2 nodes. The ends list is updated at each time step and its length corresponds to $2E(t)$. Finally, the degree list and the node list are updated. The main advantage of using a list of ends is time optimization. However, the information about the edges is lost.

1.1.2 Initial Graph

The initial number of nodes, N_0, is $m+1$, each with a degree $k_0(=m_0)=m$. These initial values meet the condition $m \leq m_0$, which is set by the original BA network [1]. This ensures that the network is complete and initially all nodes have an equal probability of being selected.

1.1.3 Type of Graph

The initial network is a simple graph with undirected edges. Upon initialisation, the network is complete.

1.1.4 Working Code

To ensure that the code is running correctly, several tests are performed:

- The ends list must have $2E(t)$ elements.
- The degree list must sum up to $2E(t)$ and its length must correspond to the number of nodes in the network.
- After adding a new node to the network, the last element of the degree list must be m, the number of nodes must increase by 1 and the length of the ends list must increase by $2m$.

1.1.5 Parameters

To initialise the `BA_model_opt` class, following parameters must be given: m, type of model - *model*, parameter for the EV model - r, and whether to use a multigraph - *multi*. For the Pure Preferential Attachment model, *model* = 'PA', *multi* = False and $m = 2, 4, 8, 16, 32, 64, 128$. These values were chosen to investigate the networks for small and big m, and also to observe the finite-size scaling effects. A new node is added by the `add_nodes` command, which needs a value of N. For this project, $N \in \{100, 1000, 10000, ...\}$ and its maximum value $N_{max} \geq m^2$.

1.2 Preferential Attachment Degree Distribution Theory

1.2.1 Theoretical Derivation

The master equation for the degree distribution is given by:

$$n(k,t+1) = n(k,t) + m\Pi(k-1,t)n(k-1,t) - m\Pi(k,t)n(k,t) + \delta_{k,m}, \tag{1.2.1}$$

where $n(k,t)$ is the number of nodes with degree k at time t, m is the number of edges added at each time step and $\Pi(k,t)$ is the probability of attaching a new edge to an existing vertex with degree k.

For the Pure Preferential Attachment model, $\Pi_{PA}(k,t) = k/(2(E(t))$. The total number of edges can be aproximated as $E(t) = mN(t)$ as $t \to \infty$. By substituting Π_{PA}, Eq. 1.2.1 can be rewritten in terms of the probability distribution $p(k,t) = n(k,t)/N(t)$:

$$p(k,t+1)N(t+1) = p(k,t)N(t) + \frac{m(k-1)}{2mN(t)}p(k-1,t)N(t) - \frac{mk}{2mN(t)}p(k,t)N(t) + \delta_{k,m}. \tag{1.2.2}$$

In the long time limit,

$$p_\infty(k) = \lim_{t\to\infty} p(k,t), \tag{1.2.3}$$

and by using

$$N(t+1) = N(t) + 1, \tag{1.2.4}$$

the probability distribution of k becomes:

$$p_\infty(k) = \frac{k-1}{2} p_\infty(k-1) - \frac{k}{2} p_\infty(k) + \delta_{k,m}. \tag{1.2.5}$$

The equation can be solved by utilising the properties of a Gamma function, Γ. If a function is given by:

$$\frac{f(x)}{f(x-1)} = \frac{x+a}{x+b} \qquad x \in \mathbb{N}, x \neq -b, \tag{1.2.6}$$

its solution is in the following form:

$$f(x) \propto \frac{\Gamma(x+1+a)}{\Gamma(x+1+b)}. \tag{1.2.7}$$

The Gamma function also possesses these properties:

$$\Gamma(x+1) = x\Gamma(x) \text{ and } \Gamma(1) = 1. \tag{1.2.8}$$

For the case $k > m$, Eq. 1.2.5 can be rearranged and solved exactly by using Eq. 1.2.7 and Eq. 1.2.8:

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k-1}{k+2} \Rightarrow p_\infty(k) = A\frac{\Gamma(k)}{\Gamma(k+3)} = \frac{A}{(k+2)(k+1)k}, \tag{1.2.9}$$

where A is the normalisation constant to be found.

For the case $k = m$ and assuming $p_\infty(k < m) = 0$, Eq. 1.2.5 becomes:

$$p_\infty(k = m) = \frac{2}{m+2} \tag{1.2.10}$$

Finally, requiring $p_\infty(k)$ to be normalised and decomposing the probability distribution in terms of partial fractions:

$$\begin{aligned} 1 =& \sum_{k=m}^{\infty} p_\infty(k) = \frac{2}{m+2} + \sum_{k=m+1}^{\infty} \frac{A}{(k+2)(k+1)k} \\ =& \frac{2}{m+2} + A\sum_{k=m+1}^{\infty}\left(\frac{1}{2(k+2)} - \frac{1}{(k+1)} + \frac{1}{2k}\right) \\ =& \frac{2}{m+2} + A\frac{1}{2(m+1)(m+2)} \end{aligned} \tag{1.2.11}$$

$$\Rightarrow A = 2m(m+1). \tag{1.2.12}$$

Hence, the exact degree probability distribution function (PDF) for the Pure Preferential Attachment model is:

$$p_\infty(k) = \frac{2m(m+1)}{(k+2)(k+1)k}. \tag{1.2.13}$$

The complementary cumulative distribution function (CCDF), $p_>(x)$, for a discrete variable, x, is defined as:

$$p_>(x) = \sum_{x'=x}^{\infty} p(x'). \tag{1.2.14}$$

For a PDF given in Eq. 1.2.13:

$$\begin{aligned} p_>(k) =& \sum_{k'=k}^{\infty} p(k') = 2m(m+1) \sum_{k'=k}^{\infty} \left(\frac{1}{2(k'+2)} - \frac{1}{(k'+1)} + \frac{1}{2k'} \right) \\ =& \frac{2m(m+1)}{2k(k+1)} = \frac{m(m+1)}{k(k+1)}. \end{aligned} \tag{1.2.15}$$

1.2.2 Theoretical Checks

The initial network consists of N_0 nodes and E_0 edges. They evolve with time as:

$$E(t) = E_0 + mt, \tag{1.2.16}$$

$$N(t) = N_0 + t. \tag{1.2.17}$$

Consider the long time limit:

$$\begin{aligned} \lim_{t\to\infty} \left(\frac{E(t)}{N(t)} \right) &= \lim_{t\to\infty} \left(\frac{E_0 + mt}{N_0 + t} \right) \\ &= \cancelto{0}{\lim_{t\to\infty} \left(\frac{E_0}{N_0 + t} \right)} + \lim_{t\to\infty} \left(\frac{m}{N_0/t + 1} \right) = m, \end{aligned} \tag{1.2.18}$$

hence the approximation $E(t) = mN(t)$ is valid. It also means that the choice of the initial network is irrelevant to the evolution of the network in the long time limit.

The probability distribution is positive for all m and k. It is also normalised as:

$$\begin{aligned} \sum_{k=m}^{\infty} p_\infty(k) =& \sum_{k=m}^{\infty} \frac{2m(m+1)}{(k+2)(k+1)k} \\ =& 2m(m+1) \sum_{k=m}^{\infty} \left(\frac{1}{2(k+2)} - \frac{1}{(k+1)} + \frac{1}{2k} \right) \\ =& 2m(m+1) \frac{1}{2m(m+1)} = 1. \end{aligned} \tag{1.2.19}$$

The complementary cumulative distribution function also obeys the requirements for $k = m$ and $k \to \infty$:

$$p_>(k = m) = \frac{m(m+1)}{m(m+1)} = 1, \tag{1.2.20}$$

$$p_>(k \to \infty) = \cancelto{0}{\lim_{k\to\infty} \frac{m(m+1)}{k(k+1)}}. \tag{1.2.21}$$

The variance of the PDF is:

$$\sigma^2 \approx \sum_{k=m}^{\infty} \frac{2m(m+1)k^2}{(k+2)(k+1)k} = \sum_{k=m}^{\infty} \frac{2m(m+1)k}{(k+2)(k+1)}, \tag{1.2.22}$$

which diverges for large k. Therefore the variance of the distribution is infinite.

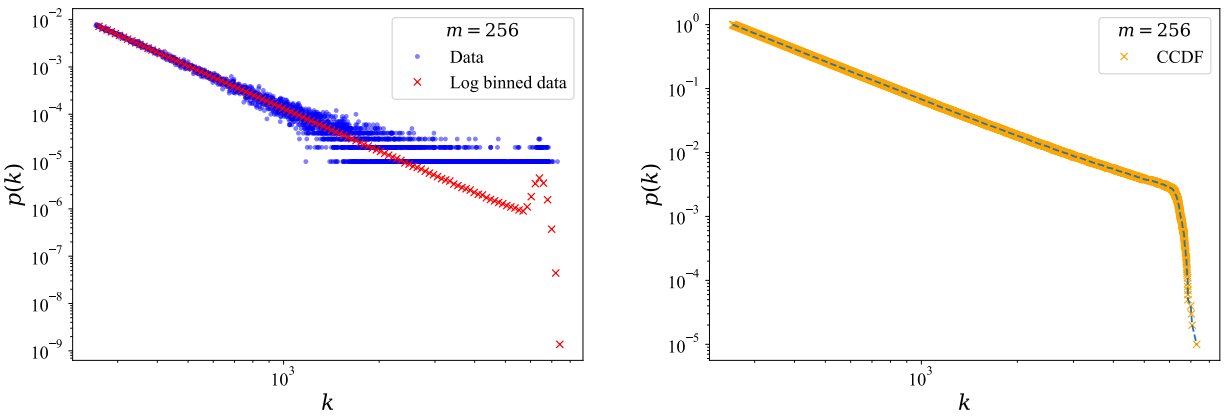
1.3 Preferential Attachment Degree Distribution Numerics

1.3.1 Fat-Tail

The degree distribution is fat-tailed as shown in Fig. 1A. It roughly follows a linear function on a doubly logarithmic plot. It also presents several issues for analysis: a few nodes have large degrees and there are values of degree which are not owned by any node. These issues are resolved in two ways:

- Log binning the data: the degrees are divided into bins of exponentially increasing width. The width of the bin j is defined as Δk^j, where Δk is a width of the first bin. For log binning, $\Delta k > 1$.
- Using the complementary cumulative distribution: the CCDF produces a smoother distribution. It is also non zero for large k, but it is still fat tailed if the PDF is fat tailed. The advantage of this method is bypassing the subjective choice of the binning.

The two methods of dealing with fat tailed distributions are shown in Fig. 1. Both smooth out the data and behave like a power-law. A characteristic bump can also be seen on the log-binned data. The bump is related to the finite-size effects, which are discussed in Section 1.3.2.

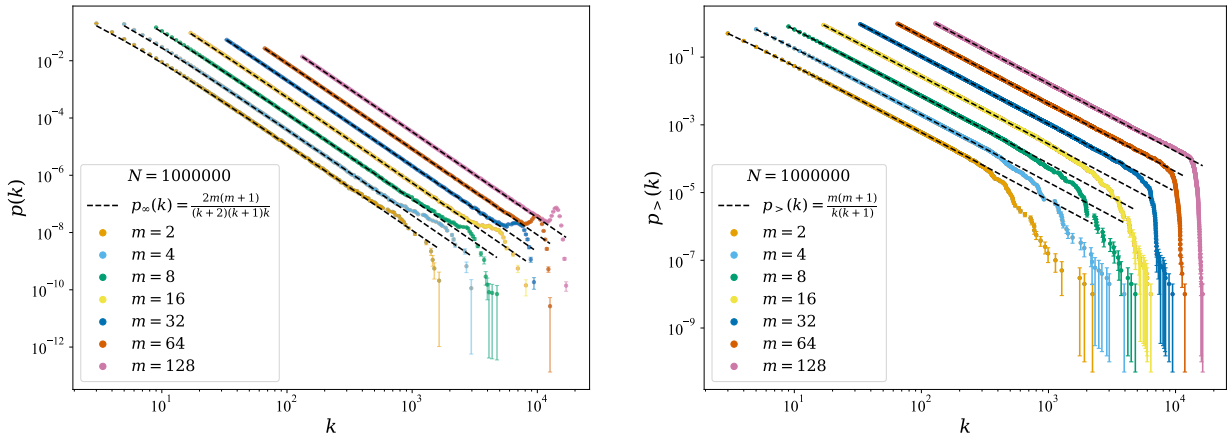


(A) (B)

Figure 1: The degree distribution for $m = 256$ and $N = 10^5$ of the PA model. A: the measured $p(k)$ is shown with blue dots. The distribution is fat-tailed and the statistical analysis of the fat-tail region is poor. The red crosses indicate the log-binned data with scale = 1.03. The distribution appears much smoother. A characteristic bump can be observed. B: the complementary cumulative distribution of the same data. The distribution is also smoothed.

1.3.2 Numerical Results

The PA model network was investigated by simulating networks with $m = 2, 4, 8, 16, 32, 64, 128$ and $N = 10^2, 10^3, .., 10^6$. Each simulation had 100 realisations to ensure good statistics. The final results were averaged across all 100 realisations and the errors were found by calculating the standard error.



(A) (B)

Figure 2: A: a log-binned degree probability distribution of the PA model acquired for varying $m = 2, 4, 8, 16, 32, 64, 128$ and fixed $N = 10^6$ with scale 1.03. The hashed black lines indicate the theoretical probability distribution. The data points are averaged over 100 realisations and the error bars are given by standard error. This is applied to all models unless stated otherwise. B: the same data expressed in terms of CCDF.

Fig. 2 shows the results for varying m and fixed $N = 10^6$. The degree distribution agrees with the theoretical probability function derived in Eq. 1.2.13. However, the PDF also displays a characteristic bump before a rapid fall. This is due to finite-size effects. The network cannot evolve infinitely, therefore the probabilities get 'accumulated' before rapidly falling off.

Fig. 3 shows the results for varying N and fixed $m = 128$. The distributions are identical, except for the characteristic bumps. The larger the size of the system, the later the bump occurs.

1.3.3 Statistics

Eq. 1.2.22 showed that the variance of this distribution is infinite. Therefore a χ^2 test cannot be performed on the distribution. Instead, a Kolmogorov-Smirnov (KS) is used and its D statistic is evaluated. It must be noted, that a KS test is only applicable to continuous distributions, whereas the degree distribution is discrete.

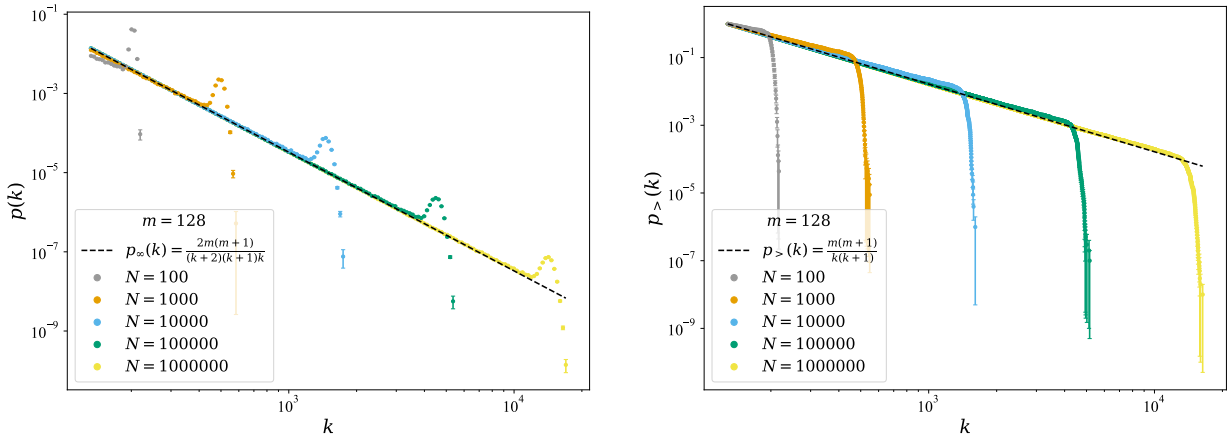
The D statistic is defined as:

$$D_n = \sup_k |F_n(k) - F(k)|, \tag{1.3.1}$$

which, in this project, can be interpreted as the largest absolute difference between the theoretical and the numerical cumulative probability distributions of k. For a perfect agreement between the numerics and theory, the D statistic must be zero.

The KS test was applied to the whole distribution, the regions before and after the bump. To ensure the analysis is valid, the regions separated by a bump are treated as separate CDFs. The results for different m and $N = 10^6$ are shown in Table 1.

Overall, the numerical results agree with the theoretical prediction as the regions before the bump have a p value of 1.000. But, the bump regions deviate from theory significantly as expected. The p value in this region is 0.000 for all m.



(A) (B)

Figure 3: A: a log-binned degree probability distribution of the PA model acquired for varying $N = 10^2, 10^3, 10^4, 10^5, 10^6$ and fixed $m = 128$ with scale 1.03. The hashed black line indicates the theoretical probability distribution. Generally, the numerical results agree with theory with the exception of bumps, which are present due to finite-size effects. B: the same data expressed in terms of CCDF.

Table 1: KS statistical test results for PA

m	2	8	32	128
Whole distribution				
D statistic	5.301×10^{-5}	4.328×10^{-05}	6.098×10^{-05}	1.876×10^{-04}
p value	0.941	0.992	0.851	0.002
Before the bump				
D statistic	5.301×10^{-5}	4.329×10^{-05}	6.102×10^{-05}	1.879×10^{-04}
p value	1.000	1.000	1.000	1.000
After the bump				
D statistic	0.165	0.104	0.036	0.054
p value	0.000	0.000	0.000	0.000

1.4 Preferential Attachment Largest Degree and Data Collapse

1.4.1 Largest Degree Theory

For a network with the total number of nodes, N, the expected number of nodes with degree larger than k_1 is equal to one:

$$\sum_{k=k_1}^{\infty} N p_\infty(k) = 1 \tag{1.4.1}$$

or

$$p_>(k_1) = \frac{1}{N}. \tag{1.4.2}$$

Substituting $k = k_1$ into Eq. 1.2.15 and rearranging yields:

$$k_1(k_1 + 1) = mN(m + 1). \tag{1.4.3}$$

This is a quadratic equation that can be exactly solved:

$$k_1 = \frac{-1 + \sqrt{1 + 4Nm(m + 1)}}{2}, \tag{1.4.4}$$

and $k_1 \sim N^{0.5}$ in the limit of large N.

The largest degree can also be approximated by considering the rate at which existing node's degree, k_i, increases as a result of a new node attaching to it:

$$\frac{dk_i}{dt} = \frac{dk_i}{dN} = m\Pi_{PA} = m\frac{k_i}{2mN}. \tag{1.4.5}$$

This leads to:

$$\frac{dk_i}{k_i} = \frac{1}{2NdN} \Rightarrow k_i = k_{i0}\sqrt{\frac{N}{N_{i0}}}. \tag{1.4.6}$$

Therefore, the largest degree, k_1, is:

$$k_1 = k_0\sqrt{\frac{N}{N_0}}, \tag{1.4.7}$$

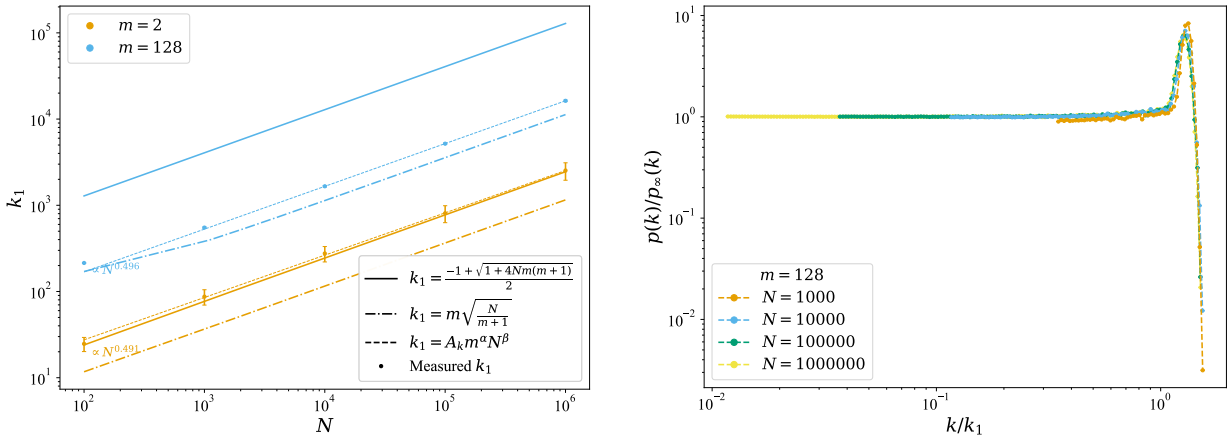
where $k_0 = m$ and $N_0 = m + 1$ based on the initial network conditions. Eq. 1.4.7 has the same dependence on N: $k_1 \sim N^{0.5}$.

1.4.2 Numerical Results for Largest Degree

The largest degree and its dependence on N was investigated. The overall shape suggested the power-law dependence, which can be expressed as:

$$k_1 = A_k m^\alpha N^\beta, \tag{1.4.8}$$

where A_k, α, and β are parameters to be fitted. For $m = 2$ and $m = 128$, the fitted β parameter was found to be $= 0.491 \pm 0.003$ and 0.496 ± 0.003, respectively. These values agree with the expected dependence. The results are shown in Fig. 4A, where the theoretical and the fitted functions are demonstrated. Clearly, Eq. 1.4.4 best describes k_1 for small m, while Eq. 1.4.7 best describes networks with larger m.



(A) (B)

Figure 4: A: the largest degree versus N for $m = 2$ and $m = 128$ of the PA model. The solid lines are the theoretical k_1 given by Eq. 1.4.4, the dash-dot lines are the theoretical k_1 given by Eq. 1.4.7, and the dashed lines are the lines of best fit assuming Eq. 1.4.8. The line of best fit shows $k_1 \propto 0.491$ and $\propto 0.496$ for $m = 2, 128$, respectively. Eq. 1.4.4 best describes k_1 for small m, while Eq. 1.4.7 best describes networks with larger m. B: a data collapse for $m = 128$ and $N = 10^3, 10^4, 10^5, 10^6$ using Eq. 1.4.7 for the expected k_1. Visually, the quality of the collapse is satisfactory, meaning this model exhibits scale-free behaviour.

1.4.3 Data Collapse

The collapse is performed by aligning the bumps. The probability distribution is divided by the theoretical prediction derived in Eq. 1.2.13. The degree values are divided by the largest degree.

The data collapse is performed on simulations with $m = 128$, therefore Eq. 1.4.7 is used to evaluate the expected k_1.

The results are shown in Figure 4B. Visually, this method of collapsing the data shows good results, where the scaled y-axis almost always is equal to 1. The bumps are also aligned due to scaling of x-axis.

2 Phase 2: Pure Random Attachment Π_{rnd}

2.1 Random Attachment Theoretical Derivations

2.1.1 Degree Distribution Theory

In the RA model, $\Pi_{RA}(k,t) = 1/N(t)$. Substituting this into Eq. 1.2.1 and taking the long time limit:

$$p_\infty(k)(m+1) = mp_\infty(k-1) + \delta_{k,m} \tag{2.1.1}$$

For cases $k > m$ and $k = m$:

$$p_\infty(k > m) = \frac{m}{m+1} p_\infty(k-1), \tag{2.1.2}$$

$$p_\infty(k = m) = \frac{1}{m+1}. \tag{2.1.3}$$

Assuming $p_\infty(k < m) = 0$, by induction:

$$p_\infty(k) = \left(\frac{m}{m+1}\right)^{k-m} p_\infty(m) \tag{2.1.4}$$

Substituting Eq. 2.1.3 yields:

$$p_\infty(k) = \left(\frac{m}{m+1}\right)^{k-m} \frac{1}{m+1}. \tag{2.1.5}$$

This PDF is positive for all m given $k \in \mathbb{N}$. It is also normalised as:

$$\begin{aligned}\sum_{k=m}^{\infty} p_\infty(k) &= \sum_{k=m}^{\infty} \left(\frac{m}{m+1}\right)^{k-m} \frac{1}{m+1} \\ &= \left(\frac{m}{m+1}\right)^{-m} \frac{1}{m+1} \left(\frac{m}{m+1}\right)^{m} (m+1) = 1.\end{aligned} \tag{2.1.6}$$

The complementary cumulative distribution function for the PDF is derived bellow:

$$\begin{aligned}p_>(k) &= \sum_{k'=k}^{\infty} p_\infty(k') = \sum_{k'=k}^{\infty} \left(\frac{m}{m+1}\right)^{k'-m} \frac{1}{m+1} \\ &= \left(\frac{m}{m+1}\right)^{-m} \frac{1}{m+1} \left(\frac{m}{m+1}\right)^{k} (m+1) = \left(\frac{m}{m+1}\right)^{k-m}.\end{aligned} \tag{2.1.7}$$

It has all of the required properties:

$$p_>(k = m) = \left(\frac{m}{m+1}\right)^{m-m} = 1 \tag{2.1.8}$$

$$p_>(k \to \infty) = \lim_{k\to\infty} \left(\frac{m}{m+1}\right)^{k-m \to 0} \tag{2.1.9}$$

The variance of the distribution can be approximated as:

$$\begin{aligned}\sigma^2 &\approx \sum_{k=m}^{\infty} k^2 \left(\frac{m}{m+1}\right)^{k-m} \frac{1}{m+1} \\ &= \left(\frac{m}{m+1}\right)^{-m} \frac{1}{m+1} \sum_{k=m}^{\infty} k^2 \left(\frac{m}{m+1}\right)^{k} = 5m^2 + m,\end{aligned} \tag{2.1.10}$$

which is finite for any network.

2.1.2 Largest Degree Theory

The largest degree, k_1, can be found via the same procedure as in Section 1.4.1:

$$p_>(k = k_1) = \frac{1}{N} \Rightarrow \left(\frac{m}{m+1}\right)^{k_1-m} = \frac{1}{N} \Rightarrow (k_1 - m) \ln \frac{m}{m+1} = \ln \frac{1}{N} \tag{2.1.11}$$

$$\Rightarrow k_1 = m - \frac{\ln N}{\ln m - \ln (m+1)}. \tag{2.1.12}$$

The largest degree can also be found by considering the rate at which node's degree increases as a result of a new node:

$$\frac{dk_i}{dN} = m\Pi_{RA} = m\frac{1}{N}. \tag{2.1.13}$$

Solving this differential equation yields:

$$k_1 = k_0 + m \ln \frac{N}{N_0}, \tag{2.1.14}$$

where $k_0 = m$ and $N_0 = m + 1$ based on the initial network conditions.

Both methods exhibit $k_1 \sim \ln N$ dependence. The RA model does not exhibit the asymptotic power-law behaviour, therefore the network is not scale-free.

2.2 Random Attachment Numerical Results

2.2.1 Degree Distribution Numerical Results

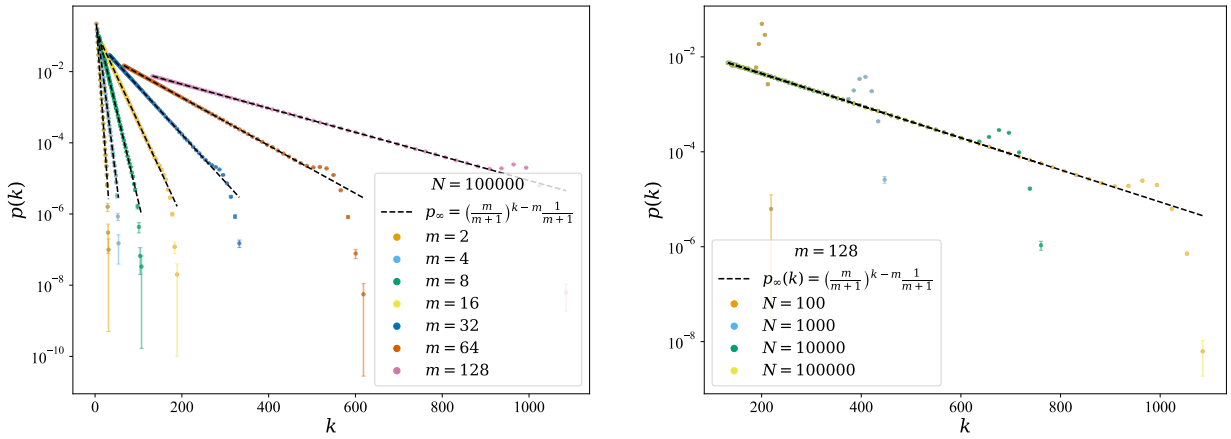
The degree distribution for the RA model is shown in Fig. 5. The uncertainty on each data point is found by calculating the standard error. The chosen m and N values are identical to the ones used in the PA model and 100 realisations were made per simulation. Visually, the numerical results agree with theory with the exception of characteristic bumps.

To further test the quality of the theoretical prediction, several tests were performed on the data. The KS test results are shown in Table 2. The p values for the region before the bump are 1.000 for all m, which indicates good agreement with the theory. The bump region does not agree with the predictions as its p values are 0.000 for all m.

Table 2: KS and χ^2 statistical test results for RA

m	2	8	32	128
Whole distribution				
D statistic	8.659×10^{-5}	1.320×10^{-04}	1.982×10^{-04}	9.199×10^{-04}
χ^2_{red}	5.394	1.408	23.782	246.745
p value	0.999	0.995	0.827	8.72779×10^{-8}
Before the bump				
D statistic	8.660×10^{-5}	1.321×10^{-04}	1.181×10^{-04}	1.221×10^{-04}
χ^2_{red}	1.193	0.747	1.245	0.792
p value	1.000	1.000	1.000	1.000
After the bump				
D statistic	5.154×10^{-2}	5.052×10^{-2}	7.534×10^{-2}	1.126×10^{-01}
χ^2_{red}	19.597	3.397	79.159	861.665
p value	0.000	0.000	0.000	0.000

An additional statistical test was also performed. Due to the probability Π_{RA} being random and the expected variance being finite for any k, a χ^2 test is applicable in this model. The reduced χ^2 statistics are shown in Table 2. For a good agreement with theory, its value must be ~ 1, which is the result obtained for the region before the bump. The region before the bump and the whole distribution agree with the theory, while the bump does not.



(A) (B)

Figure 5: A: a log-binned degree probability distribution of the RA model acquired for varying $m = 2, 4, 8, 16, 32, 64, 128$ and fixed $N = 10^5$ with scale 1.03. The hashed black lines indicate the theoretical probability distribution. Generally, the numerical results agree with theory with the exception of bumps, which are present due to finite-size effects. B: a log-binned degree probability distribution of the RA model acquired for varying $N = 10^2, 10^3, 10^4, 10^5, 10^6$ and fixed $m = 128$ with scale 1.03.

2.2.2 Largest Degree Numerical Results

The largest degree for the smallest and the biggest m were investigated and the results are shown in Fig. 6A. For small values of m both derived equations of k_1 show good agreement with numerical results, however, for $m = 128$, Eq. 2.1.14 provides a better approximation.

Moreover, a data collapse was performed on the distribution. The collapsed data for $m = 128$ is shown in Fig. 6B. The procedure is identical to the PA model collapse, and Eq. 2.1.14 was used to find expected k_1. Visually, all bumps are aligned, however small deviations are still present. This is due to RA model not following the asymptotic power-law, hence the scale-free behaviour is not expected.

3 Phase 3: Existing Vertices Model

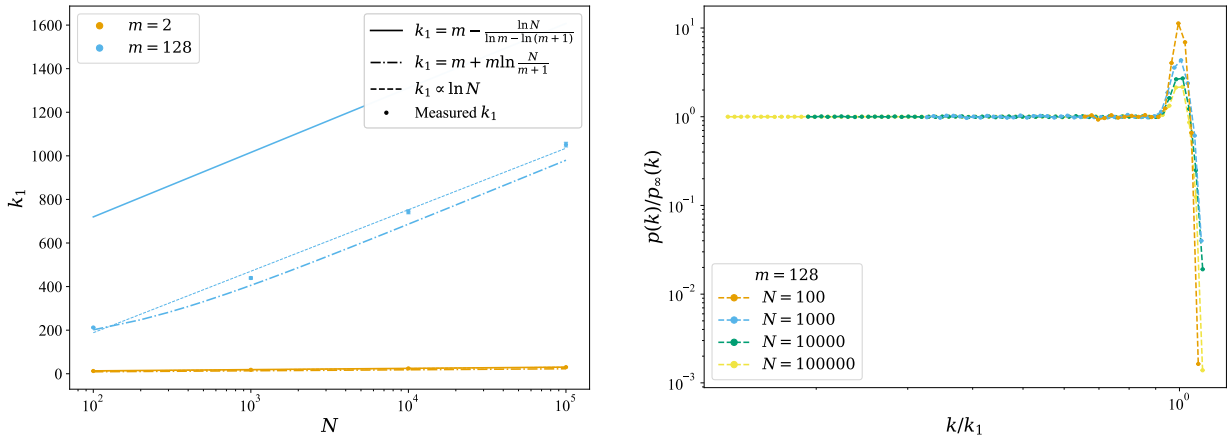
3.1 Existing Vertices Model Theoretical Derivations

Due to some of the edges being attached to the existing vertices, the Master equation given in Eq. 1.2.1 needs to be modified:

$$\begin{aligned} n(k,t+1) =& n(k,t) + r\Pi_{RA}(k-1,t)n(k-1,t) - r\Pi_{RA}(k,t)n(k,t) \\ &+ 2(m-r)\Pi_{PA}(k-1,t)n(k-1,t) - 2(m-r)\Pi_{PA}(k,t)n(k,t) + \delta_{k,r} \end{aligned} \tag{3.1.1}$$

Substituting Π_{RA} and Π_{PA}, and re-expressing the modified Master equation in terms of probability leads to:

$$p_\infty(k)\left(1 + r + \frac{(m-r)k}{m}\right) = p_\infty(k-1)\left(r + \frac{(m-r)(k-1)}{m}\right) + \delta_{k,r} \tag{3.1.2}$$



(A) (B)

Figure 6: A: the largest degree versus N for $m = 2$ and $m = 128$ of the RA model. The solid lines are the theoretical k_1 given by Eq. 2.1.12, the dash-dot lines are the theoretical k_1 given by Eq. 2.1.14, and the dashed lines are the lines of best fit assuming $k_1 \propto \ln N$. Eq. 2.1.14 best describes k_1 for larger m. B: a data collapse for $m = 128$ and $N = 10^2, 10^3, 10^4, 10^5$ using Eq. 2.1.14 for the expected k_1. Visually, the quality of the collapse is worse than the data collapse of the PA model. This is due to this model not having the asymptotic power-law behaviour.

For $k = r$, assuming $p_\infty(k < r) = 0$ Eq. 3.1.2 is rearranged:

$$p_\infty(k = r) = \frac{m}{m + 2rm - r^2}. \tag{3.1.3}$$

For $k > r$, the exact probability distribution can be obtained by using Gamma functions:

$$\frac{p_\infty(k > r)}{p_\infty(k-1)} = \frac{r + \frac{(m-r)(k-1)}{m}}{1 + r + \frac{(m-r)k}{m}} = \frac{k - 1 + \frac{mr}{m-r}}{k + \frac{m}{m-r} + \frac{mr}{m-r}} \tag{3.1.4}$$

$$\Rightarrow p_\infty(k > r) = A\frac{\Gamma(k + \frac{mr}{m-r})}{\Gamma(k + \frac{mr}{m-r} + \frac{m}{m-r} + 1)}, \tag{3.1.5}$$

where A is the normalisation constant. It can be found by utilising properties of a Beta function which is defined by:

$$B(z_1, z_2) = \int_0^1 t^{z_1 - 1}(1-t)^{z_2 - 1}dt. \tag{3.1.6}$$

It is closely related to the Gamma function:

$$B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1 + z_2)} \tag{3.1.7}$$

Changing the variables in Eq. 3.1.5 to $z_1 = k + \frac{mr}{m-r}$ and $z_2 = \frac{m}{m-r} + 1$ and using Eq. 3.1.7:

$$p_\infty(k) = A\frac{1}{\Gamma(1 + \frac{m}{m-r})}B(k + \frac{mr}{m-r}, 1 + \frac{m}{m-r}). \tag{3.1.8}$$

Forcing the probability distribution to be normalised:

$$\begin{aligned}1 =& \sum_{k=r}^{\infty} p_\infty(k) = \frac{m}{m+2rm-r^2} + \sum_{k=r+1}^{\infty} A \frac{\Gamma(k+\frac{mr}{m-r})}{\Gamma(k+\frac{mr}{m-r}+\frac{m}{m-r}+1)} \\ =& \frac{m}{m+2rm-r^2} + A\frac{1}{\Gamma(1+\frac{m}{m-r})} \sum_{k=r+1}^{\infty} B(k+\frac{mr}{m-r}, 1+\frac{m}{m-r}) \end{aligned} \tag{3.1.9}$$

Using Eq. 3.1.6 the infinite sum of the Beta function can be evaluated:

$$\begin{aligned}&\sum_{k=r+1}^{\infty} B(k+\frac{mr}{m-r}, 1+\frac{m}{m-r}) = \sum_{k=r+1}^{\infty} \int_0^1 t^{k+\frac{mr}{m-r}-1}(1-t)^{\frac{m}{m-r}+1-1} dt \\ &= \int_0^1 \sum_{k=r+1}^{\infty} t^{k+\frac{mr}{m-r}-1}(1-t)^{\frac{m}{m-r}} dt = \int_0^1 t^{\frac{mr}{m-r}+r}(1-t)^{\frac{m}{m-r}-1} dt \\ &= B(\frac{mr}{m-r}+r+1, \frac{m}{m-r}). \end{aligned} \tag{3.1.10}$$

Thus, by substutiting Eq. 3.1.10 into Eq. 3.1.9, the normalisation constant is:

$$A = \frac{2rm-r^2}{m+2rm-r^2}\Gamma\left(1+\frac{m}{m-r}\right)\frac{1}{B(\frac{mr}{m-r}+r+1, \frac{m}{m-r})}. \tag{3.1.11}$$

For simplicity, this constant will be referred to as A_{EV}. Therefore, the exact PDF for the Existing Vertices model is:

$$p_\infty(k) = A_{EV}\frac{\Gamma(k+\frac{mr}{m-r})}{\Gamma(k+\frac{mr}{m-r}+\frac{m}{m-r}+1)}. \tag{3.1.12}$$

This PDF is normalised and positive for all k.

The CCDF for this distribution is given by:

$$\begin{aligned}p_>(k) =& \sum_{k'=k}^{\infty} p_\infty(k') = A_{EV}\frac{1}{\Gamma(1+\frac{m}{m-r})}\sum_{k'=k}^{\infty} B(k'+\frac{mr}{m-r}, 1+\frac{m}{m-r}) \\ =& A_{EV}\frac{1}{\Gamma(1+\frac{m}{m-r})} B(k+\frac{mr}{m-r}, \frac{m}{m-r}) \end{aligned} \tag{3.1.13}$$

For the special case $r = m/3$, the PDF and CCDF are:

$$p_\infty(k, m=3r) = \frac{5r}{5r+3}\Gamma\left(\frac{5}{2}\right)\frac{1}{B(\frac{5r}{2}+1, \frac{3}{2})}\frac{\Gamma(k+\frac{3r}{2})}{\Gamma(k+\frac{3r}{2}+\frac{5}{2})}, \tag{3.1.14}$$

$$p_>(k, m=3r) = \frac{5r}{5r+3}\frac{1}{B(\frac{5r}{2}+1, \frac{3}{2})} B(k+\frac{3r}{2}, \frac{2}{3}). \tag{3.1.15}$$

The expected k_1 can be approximated by using the rate at which node's degree increases:

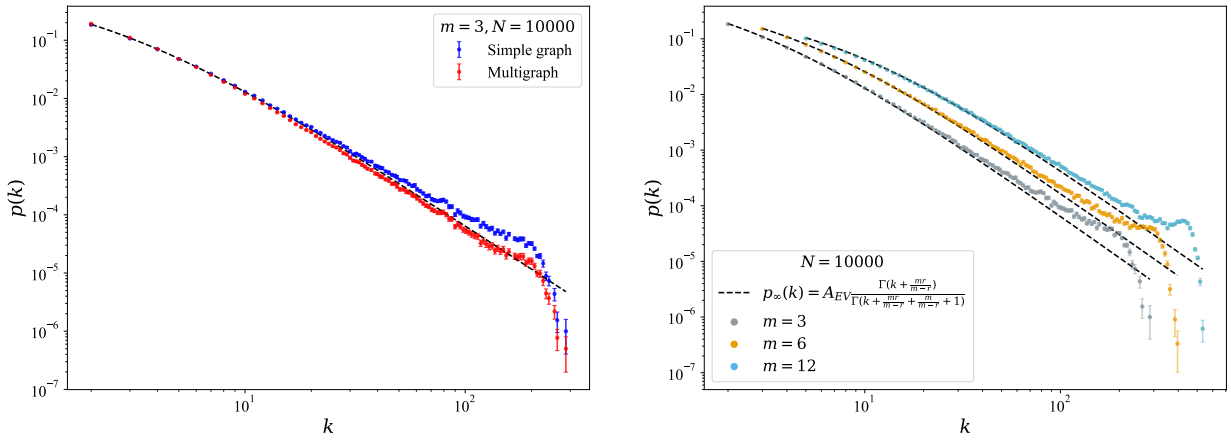
$$\frac{dk_i}{dN} = r\Pi_{RA} + 2(m-r)\Pi_{PA} = \frac{r}{N} + 2(m-r)\frac{k_i}{2mN} \tag{3.1.16}$$

Solving this differential equation leads to:

$$k_1 = \left(\frac{rm}{m-r}+k_0\right)\left(\frac{N}{N_0}\right)^{\frac{m-r}{m}} - \frac{rm}{m-r}, \tag{3.1.17}$$

where $k_0 = m$, $N_0 = 2m+1$, and $r = m/3$. The overall dependence is $k_1 \sim N^{2/3}$.

3.2 Existing Vertices Model Numerical Results



(A) (B)

Figure 7: A: a comparison of using a simple graph and a multigraph for the EV model with $m = 3$ and $N = 10^4$. The data points are averaged over 10 realisations. The multigraph follows the theory much closer, perhaps due to invalidity of assumptions made in the master equation or an insufficient number of nodes being added for the network to reach a long time limit behaviour. B: a log-binned degree probability distribution of the EV model acquired for varying $m = 3, 6, 12$ and fixed $N = 10^4$ with scale 1.03. The hashed black lines indicate the theoretical probability distribution. The numerical results poorly follow theory and the bumps due to finite-size effects are present.

Table 3: KS statistical test results for EV

m	3	6	12
Whole distribution			
D statistic	1.265×10^{-02}	1.566×10^{-02}	2.133×10^{-02}
p value	5.093×10^{-70}	5.145×10^{-107}	1.574×10^{-198}
Before the bump			
D statistic	1.271×10^{-2}	1.577×10^{-02}	2.157×10^{-02}
p value	0.833	0.364	0.014
After the bump			
D statistic	3.078×10^{-1}	3.323×10^{-1}	3.521×10^{-1}
p value	0.000	0.000	0.000

In this model, a comparison between a simple graph and a multigraph was performed and the results are shown in Fig. 7A. The multigraph results in a network which follows the theory much closer. Perhaps, this is due to assumptions in the master equation which are not valid for this model. In the master equation, it is assumed that a nodes degree can only increase by 1 at a time. But, in this model, a node's degree can increase by a value up to $m - r$.

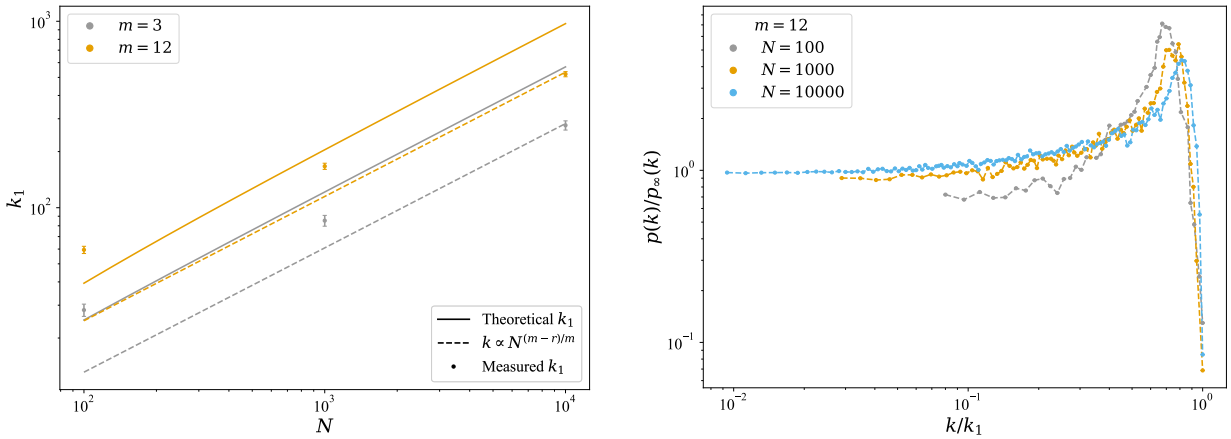
For the purposes of consistency, simple graphs were only investigated. The degree distributions averaged over 10 realisations for $m = 3, 6, 12$ and $N = 10000$ are shown in

Fig. 7B. Visually, the numerical results poorly follow the theory, therefore poor statistical results are expected.

A KS test was performed and the corresponding results are shown in Table 3. The p value for larger m is less than 0.05 even for a region before the bump, indicating that the numerical results do not follow theory. Perhaps, this is due to an insufficient number of nodes being added for the network to reach the long time limit.

The largest degree distribution for the EV model is shown in Fig. 8A. The numerical results do not follow the expected N dependency.

The collapse of this model is shown in Fig. 8B. For this collapse, the measured k_1 is used. The quality of this collapse is much worse comapred to the other two models.



(A) (B)

Figure 8: A: the largest degree versus N for $m = 3$ and $m = 12$ of the EV model. The measured values are averaged over 10 realisations. The solid lines are the theoretical k_1 given by Eq. 3.1.17 and the dashed lines are the lines of best fit assuming $k_1 \propto N^{(m-r)/m}$. The numerical results do not agree with theory. B: a data collapse for $m = 12$ and $N = 10^2, 10^3, 10^4$ using measured k_1. Visually, the quality of the collapse is the worst compared to the other two models.

4 Conclusions

The PA and RA models showed good consistency with theory and a good quality of the data collapse. This was validated by statistical tests, and the p values for the region before the bump were 1.000 for all m. The EV model showed the worst agreement with theory. Even the regions before the bump had a p value less than 0.05 for larger m. More investigation is needed for this model.

References

[1] A.-L. Barabási, R. Albert, *Emergence of scaling in random networks*, Science 286 (1999) 509–512. Available from: https://link.aps.org/doi/10.1103/PhysRevLett.59.381 [Accessed: March 2023]