Network Project
A Growing Network Model

CID: 01807529

27th March 2020

Abstract: The aim of the project is to study the Barabási and Albert model. This model has been devised to reproduce a common property observed in many real-world networks, namely a scale-free power-law distribution of the connectivity of the nodes. The model consists of a growing network where new nodes prefer to attach to the more connected nodes. The preferential attachment is compared with two other growing network models for which the choice of the target network is made randomly or by means of a random walk.

Word Count: 2800

1 Introduction

A high degree of self-organization characterizes the large-scale properties of complex networks such as the World Wide Web or the citation network of scientific papers. The Barabási and Albert model aims to incorporate two fundamental features of real-world networks: growth and preferential attachment. Random network models that fail to incorporate these two features also fail to give the power-law scaling observed in real networks. In the first part of the project, the Barabási and Albert model is introduced theoretically. The degree distribution is compared with that obtained with a numerical simulation by means of statistical tests. In the same framework, the preferential attachment is substituted with a purely random choice of the target node for the edges to be added. Lastly, a third growing network is studied. In this model, the choice of the target node in the growing process is made by means of a random walk.

1.1 Definition

In order to define the Barabási and Albert (BA) model, let us start with the definition of a *random graph*.

Definition 1. *A **random graph** is a set of graphs $\mathcal{G} = \{G_1, G_2, G_3, \dots\}$ on which a probability measure $\mathbb{P}$ is assigned such that $\mathbb{P}(G_i) \in [0, 1]$.*

The BA model studied in this report is a succession of simple, undirected random graphs:

Definition 2. *The **Barabási and Albert (BA) model** is a succession of simple, undirected random graphs $\{\mathcal{G}_{t=0,1,\dots,\infty}\}$ with the following properties:*

- $\mathcal{G}_0 = \{G_0\}$ *with* $|N(G_0)| = m_0$

- $|N(G_{t+1})| = |N(G_t)| + 1$, $\forall G_t \in \mathcal{G}_t$ *and* $G_{t+1} \in \mathcal{G}_t$
- $|E(G_{t+1})| = |E(G_t)| + m$, $\forall G_t \in \mathcal{G}_t$ *and* $G_{t+1} \in \mathcal{G}_{t+1}$, $m \leq m_0$.

Where $|N(\dots)|$ and $|E(\dots)|$ are the cardinalities of the set of vertices and edges respectively. Let $n(k,t)$ be the number of vertices in the graph with degree k: $n(k,t)$ is then a succession of random variables.

Definition 3. *With $n(k,t)$ being a succession of random variables, the* ***conditional expected value*** *is $\mathbb{E}(n(k,t+1)|G_t)$, with $\mathbb{E}(n(k,t)|G_t) = n(k,t)$.*

The model is compleatly determined by assigning a transition probability:

Definition 4. *In this model the* ***transition probability*** *is assigned as: $\mathbb{P}(G_{t+1}|G_t) \Leftrightarrow \mathbb{P}((t+1,s) \in G_{t+1}|G_t)$, $s \in E(G_t)$.*

In the BA model the transition probability is

$$\mathbb{P}((t+1,s) \in G_{t+1}|G_t) := \Pi_{\text{pa}}(k_s,t) := \frac{k_s}{\sum_{s'} k_{s'}} = \frac{k_s}{2|E(G_t)|} \tag{1}$$

2 Phase 1: Pure Preferential Attachment Π_{pa}

2.1 Implementation

2.1.1 Numerical Implementation

In the programme a graph is defined by a list of lists Nb = [[...], [...], [...], ...] which is the list of neighbours of each vertex: the i^{th} element of Nb is the list of vertices linked to the vertex labeled i. In order to implement preferential attachment, a second list "a" keeps count of all the stubs of each vertex. Whenever two vertices i and j are linked together, their label is added to "a" so that by randomly choosing an entry from this list, the probability of choosing a particular vertex is proportional to its degree:

$$\begin{aligned} &\text{let } j \in [0, |Nb| - 1] \text{ be a vertex} \\ &\text{let } i \in [0, |a| - 1] \text{ be a randomly choosen index} \\ &\mathbb{P}(a[i] = j) = \frac{|Nb[j]|}{|a|} = \frac{k_j}{\sum_s k_s}. \end{aligned} \tag{2}$$

Figure 1 shows an example of a graph represented by:

$$\begin{aligned} Nb &= [[1,2,3,4],[0,2,3,4],[0,1,3,4],[0,1,2,4],[0,1,2,3]] \\ a &= [0,0,0,0,1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4] \end{aligned}$$

2.1.2 Initial Graph

In Section 2.2.1 and 3.1.1, the exact form for the degree distributions in the long-time limit is derived. Apart from the long-time limit approximation, for both preferential attachment and random attachment, the theoretical distribution is more accurate if the initial graph is such that $E(0) = mN(0)$ as described in Eq. (5). The graph with the smallest N that satisfies this requirement is the complete graph with $N = 2m + 1$. An example of such a graph with $m = 2$ is the one in Fig. 1.

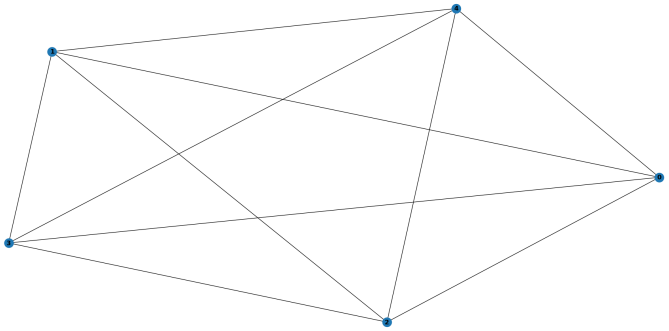


Figure 1: Simple undirected network with $N = 5$ and $E = 10$. The network is also complete.

2.1.3 Type of Graph

The BA model has been devised in order to describe real-world systems that seem to self-organize into a scale-free stationary state. Examples of these networks are the WWW (World Wide Web) or the citation network. The former is the network of web pages (nodes) that point to other web pages via hyperlinks (edges). The latter is the citation pattern of the scientific publications, where nodes are the published papers and edges are the links to the papers cited. In both cases, the relation represented by the edges is not symmetric, so the graph representing the WWW and the citation network would be directed. In this report, though, we deal with undirected graphs. It turns out that the distribution of both the *indegree* of a directed graph and the degree of simple graphs resulting from preferential attachment follow a power-law decay.

In the paper published in 1999 by Barabási and Albert the model is described in this way:

"To incorporate the growing character of the network, starting with a small number (m_0) of vertices, at every time step we add a new vertex with $m(\leq m_0)$ edges that link the new vertex to m different vertices already present in the system. " [1]

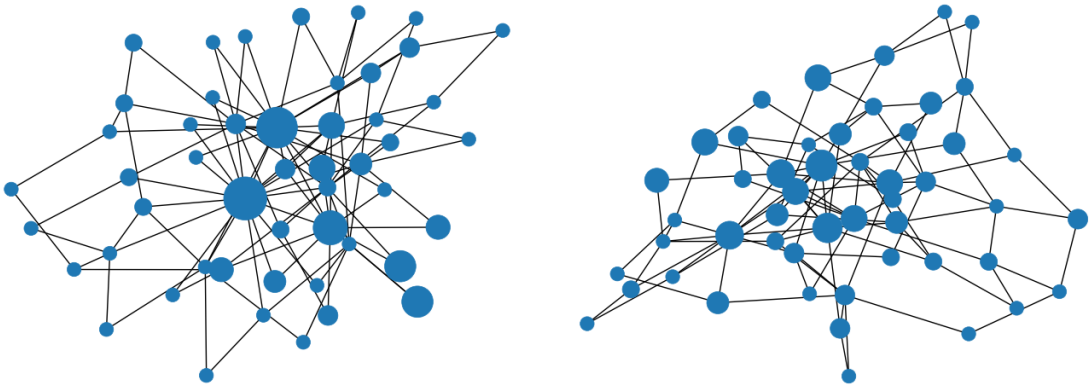
In conclusion, all the models in this report generate simple undirected graphs with neither self-loops nor multiple edges between vertices. These models differ from the one described in the quote for the fact that the initial graph always has $mN(0)$ edges.

2.1.4 Working Code

A qualitative analysis of the representation of the network for small N is a good starting point to test whether the programme is working correctly: figure 2 shows two networks realized using preferential attachment and random attachment.

A quantitative test has been performed by looking at the boundary value of $p(k = m)$ for different m and different N. Table 1 and 2 report the results together with the expected values. The errors have been estimated by $\frac{\sigma}{\sqrt{M}}$ where M is the number of realizations of the system, and σ is the standard deviation.

Lastly, it was verified that the average degree, namely $K = \frac{\sum_{i=0}^{N} Nb[i]}{N}$, was equal to $2m$ for all N.



(a) *Preferential Attachment* (b) *Random Attachment*

Figure 2: Representaion of two networks generated with the two different algorithms. The dimension of the nodes is proportional to their connectivity

N	$m = 1$	Exp.	$m = 3$	Exp.
10^4	0.6672 ± 0.0002	0.66667	0.4013 ± 0.0005	0.4
10^5	0.6668 ± 0.0001	0.66667	0.3996 ± 0.00007	0.4
10^6	0.66697 ± 0.00007	0.66667	0.39997 ± 0.00004	0.4

Table 1: Preferential attachment: numerical and expected values of $p(k = m)$ for different N and different m. The errors have been estimated by $\frac{\sigma}{\sqrt{M}}$ where M is the number of realizations of the system.

N	$m = 1$	Exp.	$m = 3$	Exp.
10^4	0.5002 ± 0.0002	0.5	0.2501 ± 0.0002	0.25
10^5	0.5003 ± 0.0001	0.5	0.2497 ± 0.0001	0.25
10^6	0.50001 ± 0.00002	0.5	0.25014 ± 0.00003	0.25

Table 2: Random attachment: numerical and expected values of $p(k = m)$ for different N and different m. The errors have been estimated by $\frac{\sigma}{\sqrt{M}}$ where M is the number of realization of the system.

2.1.5 Parameters

The parameters required by the programme are the number of nodes N to be added the number m of edges to be added every time step. For some tasks, the system is realized M times using the same parameters, and the result is the average over the different realizations. The last part of the programme that implements the random walk requires the value of q, namely the probability of ending the walk after changing vertex.

2.2 Preferential Attachment Degree Distribution Theory

2.2.1 Theoretical Derivation

Using the notation of section 1.1 let's consider the conditional expected value of $n(k,t)$:

$$\mathbb{E}(n(k,t+1)|G_t) = n(k,t) + m\Pi_{\text{pa}}(k-1,t)n(k-1,t) - m\Pi_{\text{pa}}(k,t)n(k,t) + \delta_{k,m} \tag{3}$$

From Definition 2. we have:

$$\begin{aligned} E(t) &= m_0 + t \\ N(t) &= E(0) + mt \end{aligned} \tag{4}$$

which implies

$$\begin{aligned} &E(t) \cong mN(t) \text{ and} \\ &E(t) = mN(t) \text{ if } E(0) = mN(0). \end{aligned} \tag{5}$$

Where the shorthand notation $|E(G_t)| \equiv E(t)$ and $|N(G_t)| \equiv N(t)$ has been used. We can write Eq. (3) in terms of the probability distribution $p(k,t) = \frac{n(k,t)}{N(t)}$ and look for solutions that are stable in the long time limit $t \to \infty$, which means:

$$p(k,t) = \frac{n(k,t)}{N(t)} = \frac{n(k,t+1)}{N(t+1)} = p(k,t+1) := p_\infty(k). \tag{6}$$

We have:

$$\begin{aligned} \mathbb{E}(n(k,t+1)|G_t) &= n(k,t) + m\Pi_{\text{pa}}(k-1,t)n(k-1,t) - m\Pi_{\text{pa}}(k,t)n(k,t) + \delta_{k,m} = \\ &= n(k,t) + m\frac{(k-1)n(k-1,t)}{2mN(t)} - m\frac{kn(k,t)}{2mN(t)} + \delta_{k,m}, \end{aligned} \tag{7}$$

that expressed in terms of $p(k,t)$ becomes:

$$\begin{aligned} &\mathbb{E}(N(t+1)p(k,t+1)|G_t) = (N(t)+1)\mathbb{E}(p(k,t+1)|G_t) = \\ &= N(t)p(k,t) + \frac{1}{2}(k-1)p(k,t) - \frac{1}{2}kp(k,t) + \delta_{k,m} \overset{t\to\infty}{=} \\ &\overset{t\to\infty}{=} N(t)p_\infty(k) + \frac{1}{2}(k-1)p_\infty(k) - \frac{1}{2}kp_\infty(k) + \delta_{k,m} \end{aligned} \tag{8}$$

If the stationary solution exists, then it is compleatly determined by the initial conditions and we have

$$\mathbb{E}(p_\infty(k)|G_t) = p_\infty(k). \tag{9}$$

The master equation for the BA model is therefore

$$p_\infty(k) = \frac{1}{2}(k-1)p_\infty(k-1) - \frac{1}{2}kp(k) + \delta_{m,k}, \text{ for } k \geq m. \tag{10}$$

To find the solution to this recursion formula we first consider the case $k > m$. By rearrangeing it as

$$\frac{p_\infty(k)}{p_\infty(k-1)} = \frac{k-1}{k+2}, \tag{11}$$

we can show that

$$p_\infty(k) = A\frac{\Gamma(k)}{\Gamma(k+3)} = A\frac{(k-1)!}{(k+2)!} = \frac{A}{k(k+1)(k+2)} \tag{12}$$

is a solution. The coefficient A can be determined by cosidering the boundary case $k = m$ in Eq. (10):

$$\begin{aligned} p_\infty(m) &= \frac{1}{2}(m-1)p(m-1) - \frac{1}{2}mp(m) + \delta_{m,m} = \\ &= \frac{1}{2}mp(m) + 1 \Rightarrow p_\infty(m) = \frac{2}{2+m} \end{aligned} \tag{13}$$

where the fact that $p_\infty(k < m) = 0$ has been used: we can chose an initial graph with $E(0) = mN(0)$ so that, by adding m edges every step, there will not be any vertex with $k < m$. By imposing the boudary condition in Eq. (12) we find the exact degree distribution $p_\infty^{pa}(k)$ in the long-time limit for preferential attachment in the BA model:

$$p_\infty^{pa}(k) = \frac{2m(m+1)}{k(k+1)(k+2)}, \text{ for } k \geq m \tag{14}$$

2.2.2 Theoretical Checks

The first theoretical check for the probability distribution is that it is normalized:

$$\begin{aligned} &\sum_{k=m}^{\infty} n(k) = \sum_{k=m}^{\infty} \frac{2m(m+1)}{k(k+1)(k+2)} = \\ &= 2m(m+1) \lim_{L\to\infty} \left[\frac{1}{2} \sum_{k=m}^{L} \left(\frac{1}{k} - \frac{1}{(k+1)} \right) + \frac{1}{2} \sum_{k=m}^{L} \left(\frac{1}{k+2} - \frac{1}{k+1} \right) \right] = \\ &= \frac{m(m+1)}{m(m+1)} = 1. \end{aligned} \tag{15}$$

Secondly, it's clear that, as expected, the decay follows a power-law that for large k is $\propto k^{-3}$.

2.3 Preferential Attachment Degree Distribution Numerics

2.3.1 Fat-Tail

The degree distribution $n(k)$ of a BA model network with $N = 10^6$ and $m = 2$ has non-zero values only for $k \geq k_c$ with $k_c \approx 150$. Above this value there are only a few vertices with $n(k > k_c) > 0$ and many points with $n(k > k_c) = 0$ that must be ignored when plotting $n(k)$ vs. k in a log-log plot. Figures 3, 4 and 5 show three different ways to deal with the noisy tail of the distribution, and improve the data visualization; namely the data binning, the cumulative distribution and the Zipf plot. By log-binning the data, we can extract information from the tail of the observed distribution. We divide the k-axis into bins of exponentially increasing length: the j^{th} bin will cover the interval $[a^j, a^{j+1}[$, with $a > 1$. The probability distribution is then defined as

$$\tilde{n}_N(k^j) = \frac{\text{Number of nodes with } k \in [k_{min}^j, k_{max}^j]}{N\Delta k^j}, \tag{16}$$

where k_{min}^j, k_{max}^j are the mininimum and maximum integers in bin j and

$$\begin{aligned} \Delta k^j &= (k_{max}^j - k_{min}^j + 1) \text{ is the number of integers in the interval,} \\ k^j &= \sqrt{k_{max}^j k_{min}^j}. \end{aligned} \tag{17}$$

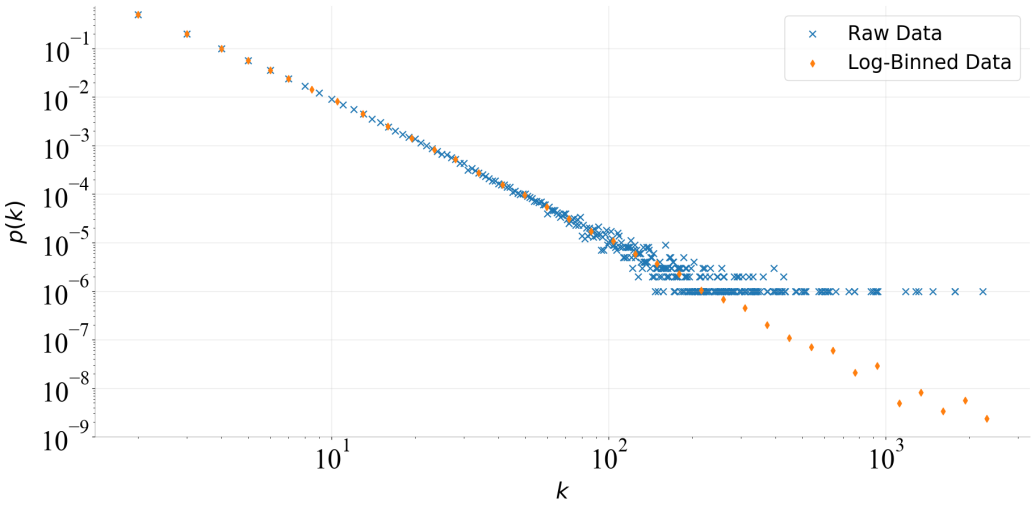


Figure 3: Log-log plot of the degree distribution of a BA model network with $N = 10^6$ vertices, average degree $K = 4$ and $m = 2$, together together with the log binned data. The log binning has been performed using $a = 1.2$.

The cumulative distribution is defined as

$$p_<(k) = \sum_{k'=0}^{k} n(k'). \tag{18}$$

The Zipf plot shows the degree k vs. the rank r of the vertex with the r^{th} largest degree.

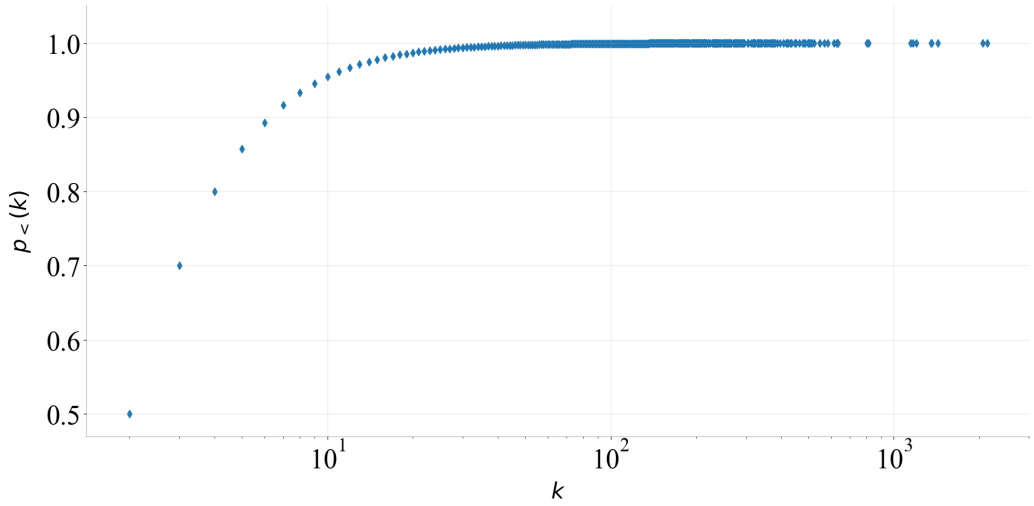


Figure 4: Cumulative distribution function (CDF) of a BA model network with $N = 10^6$ vertices, average degree $K = 4$ and $m = 2$.

2.3.2 Numerical Results

The numerical data for fixed $N = 10^6$ and $m = 1, 2, 3, 4, 10$ are compared with the theoretical results in Fig. 6. The numerical data shown in the figure have been obtained

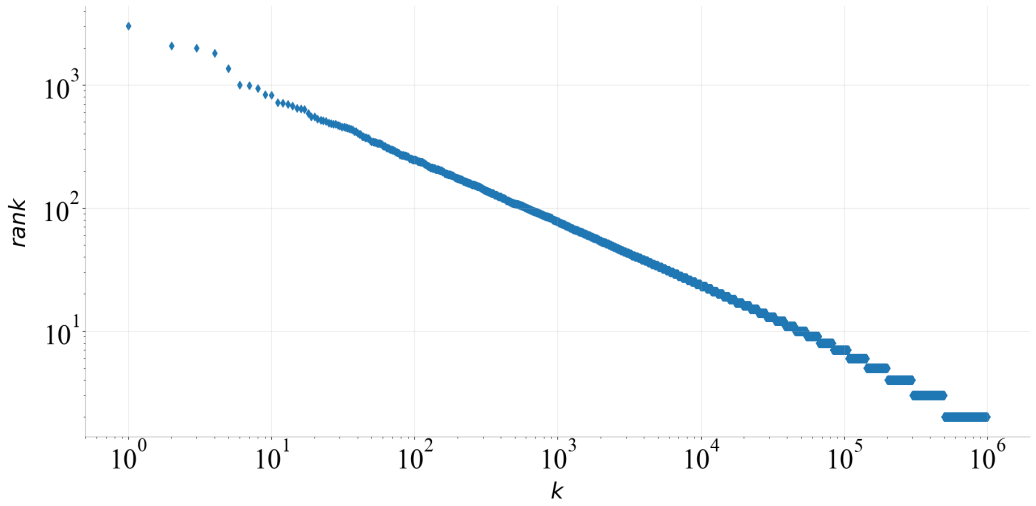


Figure 5: Degree k vs. *rank* of a BA model network with $N = 10^6$ vertices, average degree $K = 4$ and $m = 2$.

by averaging over 100 different realizations and by log-binning, as described in Eq. (16), using $a = 1.1$. The dashed line represents the theoretical form of the distribution; note that it has been represented as a continuous line only to make more clear whether the numerical data fit the theoretical result. The theoretical distribution is indeed discrete, and the lines connecting its points are meaningless. A qualitative analysis suggests a good

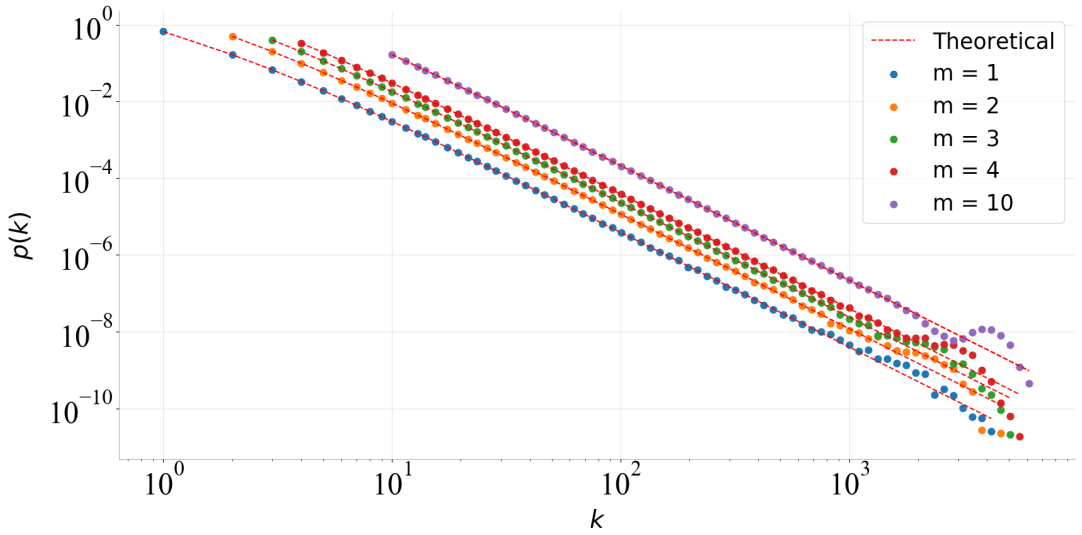


Figure 6: Preferential attachment: numerical result, in a log-log plot, of the degree distribution for fixed $N = 10^{10}$ and $m = 1, 2, 3, 4, 10$, together with the theoretical results. The numerical data is the result of $M = 100$ realizations of the system and a log-binning process with $a = 1.1$.

fit of the numerical data to the theoretical results for small values of k. For large values of k finite-size effects make the numerical data deviate from the theoretical distribution that has been obtained in the long-time limit and therefore is valid for systems with a number of nodes $N \to \infty$. The finite-size effects become more evident as m icreases. This has not been investigated properly, but it can be traced back to the fact that the degree of a

vertex in a network with N vertices is bounded by $N-1$, which in some sense represents a cut-off degree that forces the distribution to zero.

2.3.3 Statistics

The statistical analysis performed on the numerical data consists of two statistical hypothesis tests: the χ^2 test and the Kolmogorov-Smirnov (KS) test. For the χ^2 test, we define the variable

$$\sum_j \frac{(o_j - e_j)^2}{e_j} \tag{19}$$

with o_j and e_j being the observed and expected frequencies, respectively. Let's consider a particular class k_j. Let $P(k_j)$ be the probability of the event $k = k_j$ so that $1 - P(k_j)$ is the probability of the event $k \neq k_j$. e_j would follow a binomial distribution with expected value $NP(k_j)$. If the expected frequency is large enough, in particular, if it is larger then 5, then the binomial distribution tends to a Poissonian distribution with expected value $NP(k_j)$ and standard deviation $\sqrt{NP(k_j)}$. The Poissonian distribution tends in turn to a Gaussian distribution with the same parameters so that the (19) becomes

$$\sum_j \frac{(o_j - NP(k_j)^2}{(NP(k_j))^2} = \sum_j \frac{(o_j - NP(k_j))^2}{\sigma_j^2} = \chi^2 \tag{20}$$

which is the well known chi-squared variable. We make the hypothesis H_0 that the numerical distribution can be described by the theoretical one. By evaluating $P(\chi^2 \geq \chi_0^2)$, namely the probability of finding a χ^2 larger then the observed χ_0^2, we either accept H_0 or reject it with a fixed significance level α. In practice, the observed frequencies are larger than 5 only for small values of k so that the test has been performed by evaluating the variable

$$\chi_0^2 = \sum_{k=1}^{k_c - 1} \frac{(n_o(k) - n_e(k))^2}{n_e(k)} \tag{21}$$

where k_c is the smallest value of k such that $n_o(k_c) < 5$ and $n_e(k) = Np_\infty(k)$. The test has been performed for $m = 1, 2, 3, 4, 5$ and the results are reported in Table 3. Figure 7 shows the numerical and theoretical distributions used in the test. The one-sample KS test is a nonparametric test of the equality of continuous probability distributions. It is based on the distance between the numerical CDF of the sample and the CDF of the theoretical distribution. Since the theoretical CDF is not continuous, the two-sample KS test has been implemented instead. The two-sample KS statistic is

$$D = \sup_k |F_1(k) - F_2(k)|, \tag{22}$$

where $F_{1,2}(k)$ are the CDFs. The hypothesis H_0 is that the two CDFs come from the same distribution. We either accept or reject H_0 with a fixed significance level based on the value of $P(D \geq D_0)$. In practice, the test has been performed using the numerical CDF as F_1, and as F_2 the CDF of a sample of variables drown from a population distributed as

$$p_{ks}(k) = \begin{cases} p_\infty(k) & k \leq k_c \\ 1 - \sum_{k=1}^{k_c} p_\infty(k) & k > k_c \end{cases}$$

The results have been reported in Table 3.

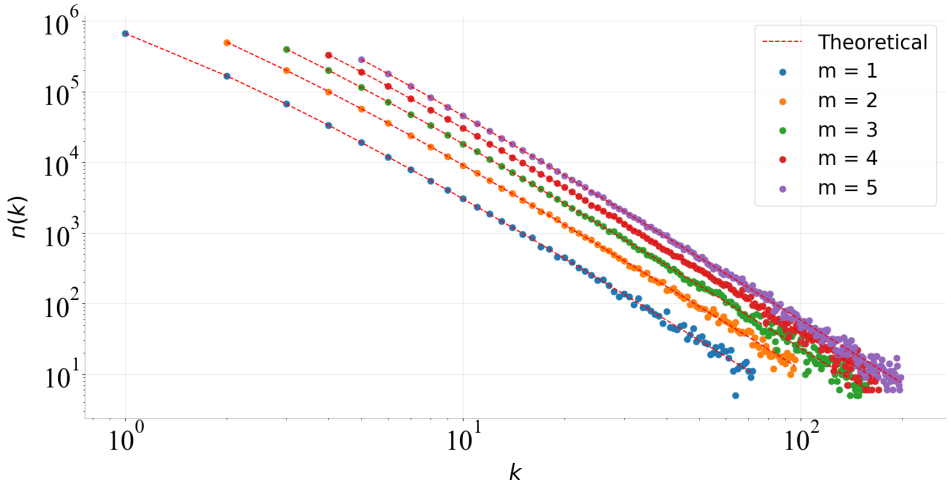


Figure 7: Preferential attachment: log-log plot of the frequency distributions used for the χ^2-test. The frequency needs to be larger than 5.

m	χ^2	p_{χ^2}	D	p_D
1	63.8	0.716	0.0006	0.993
2	60.19	0.997	0.0003	1
3	143.5	0.697	0.0008	0.934
4	161.8	0.578	0.008	0.895
5	184.6	0.636	0.00119	0.477

Table 3: Results of the statistical tests for preferential attachment. The null hypothesis are that the numerical distributions can be described by the theoretical one (for χ^2-test) and that the two samples come from the same population (KS test). For both we can accept H_0 with a high significance level.

2.4 Preferential Attachment Largest Degree and Data Collapse

2.4.1 Largest Degree Theory

Let us start by analyzing the behaviour of the single nodes as the graph grows. In order to do so, let's use a continuous real variable to represent the average value of the degree k_s of the s^{th} added node. We set up a differential equation for this variable, and we solve it in the long-time limit.

$$\frac{dk_s}{dt} = m\Pi_{\text{pa}}(k_s, t) = m\frac{k_s}{2mt - m} \overset{t>>1}{=} \frac{k_s}{2t}. \tag{23}$$

The solution of this differential equation is

$$k_s(t) = m_0 \left(\frac{t}{t_0}\right)^{\frac{1}{2}} \tag{24}$$

where time is to be interpreted as the "event-time", namely a continuous parameter that advances proportionally to the arrival of a new node in the graph. t_0 is the time at which

the node is added to the graph and $k_s(0) = m_0$. We expect the largest degree in the BA model to grow sublinearly as

$$k_1 \propto N^{\frac{1}{2}} \tag{25}$$

By looking at the expression of the rank of a vertex k we find

$$\begin{aligned} r(k) &= \sum_{k'=k}^{\infty} n(k') = \sum_{k'=k}^{\infty} N p_{\infty}^{\mathrm{pa}}(k) = \\ &= \sum_{k'=k}^{\infty} \frac{N 2m(m+1)}{k(k+1)(k+2)} = \\ &= N 2m(m+1) \lim_{L\to\infty} \frac{1}{2} \sum_{k'=k}^{L} \left(\frac{1}{k'} - \frac{1}{(k'+1)} \right) + \frac{1}{2} \sum_{k'=k}^{L} \left(\frac{1}{k'+2} - \frac{1}{k'+1} \right) = \\ &= \frac{Nm(m+1)}{k(k+1)}. \end{aligned} \tag{26}$$

By solving for k_1, namely the degree of the vertex with rank 1, we find:

$$k_1^2 + k_1 - Nm(m+1) = 0 \Rightarrow k_1 = \frac{-1 \pm \sqrt{1 + 4Nm(m+1)}}{2} \tag{27}$$

The largest degree in a finite system depends on N in the form of Eq. (25). More precisely, for large N:

$$k_1(N) = \sqrt{m(m+1)N}. \tag{28}$$

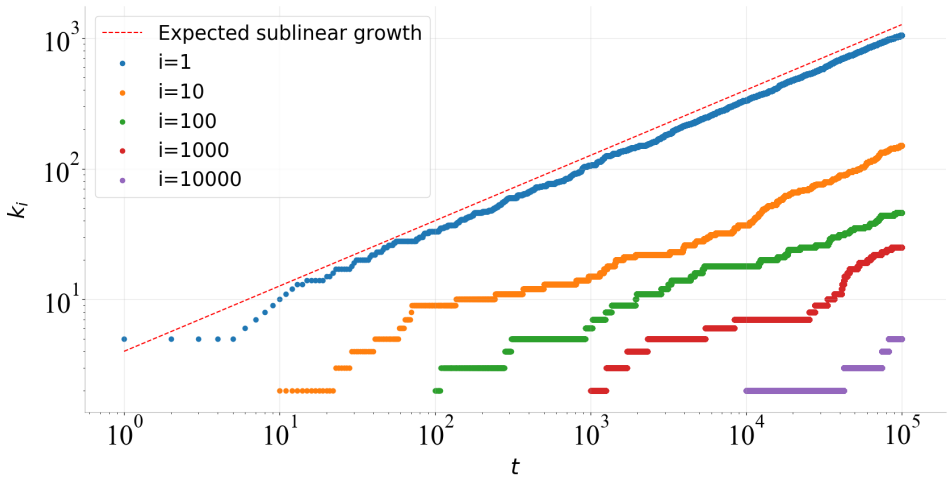


Figure 8: Preferential attachment: log-log plot of k_i vs N for $i = 1, 10, 100, 1000, 10000$, with k_i the degree of the vertex labeled i. For $\leq 2m + 1$, the vertex is part of the initial graph. The dashed line shows the expected sublinear behaviour.

2.4.2 Numerical Results for Largest Degree

Figure 8 shows a log-log plot of k_i vs N for $i = 1, 10, 100, 1000, 10000$, with k_i the degree of the vertex labeled i. Note that for $i \leq 2m + 1$, the vertex is part of the initial graph.

In the same figure, the dashed line shows the expected sublinear behaviour of Eq. (24). Since $m = 2$ gives the best results in the statistical tests, this value has been used for the next numerical simulations. The study of the behaviour of k_1 and N consists of a linear regression of the set of points $\{(\log_{10} N, \log_{10} k_1)$, for N $= 10^2, 10^3, 10^4, 10^5, 10^6\}$. Each degree is the result of the average over $M = 100$ realizations, so the errors have been estimated as $\frac{\sigma}{\sqrt{M}}$, with being the σ the standard deviation. Figure 9 shows the data with the error bars, together with the regression line and the theoretical behaviour of Eq. (28). Table 4 summarizes the result of the fit and reports errors that are too small to be appreciated in the figure together with the result of a χ^2-test of the goodness of the fit performed using the variable

$$\chi^2 = \sum_{i=2}^{6} \frac{\left(y_i^{num} - y_i^{fit}\right)^2}{\widetilde{\sigma}_i^{\,2}}, \tag{29}$$

where $\widetilde{\sigma}_i = \frac{\sigma_i}{y_i}$, is the error to be used when we linearize $y = ax^b$.

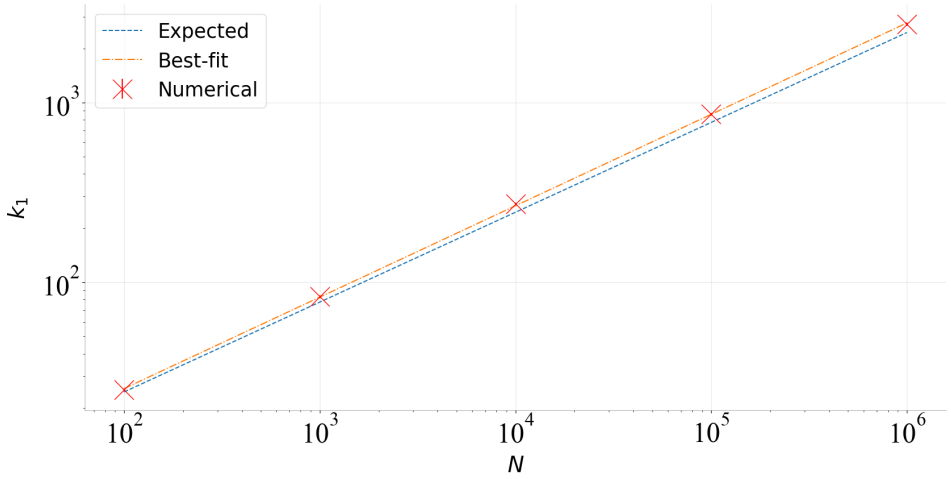


Figure 9: Preferential attachment: log-log plot of the largest degree data with the error bars, together with the regression line and the expected behavior: $\log_{10} k_1 \propto 0.5 \log_{10} N$.

σ_1	σ_2	σ_3	σ_4	σ_5	χ^2	p_{χ^2}	c_1	c_0
0.45	1.53	4.25	16.68	53.77	0.9	0.99	0.509 ± 0.004	0.40 ± 0.02

Table 4: Errors and results of the χ^2-test for the largest degree in preferential attachment. The last two columns are the parameters of the regression line $c_1 \log_{10}(N) + c_0$. The expected values are $c_1 = 0.5$ and $c_0 = 0.38$.

2.4.3 Data Collapse

Using the scaling relation of k_1 and the theoretical form of the probability distribution, we can vertically align the rapid decay (due to finite size effects) of each graph by plotting (in a log-log plot) the transformed probability $p(k)/p_\infty(k)$ vs. k. Using the rescaled variable k/k_1 the decays are aligned horizontally. The data collapse is shown in Fig. 11.

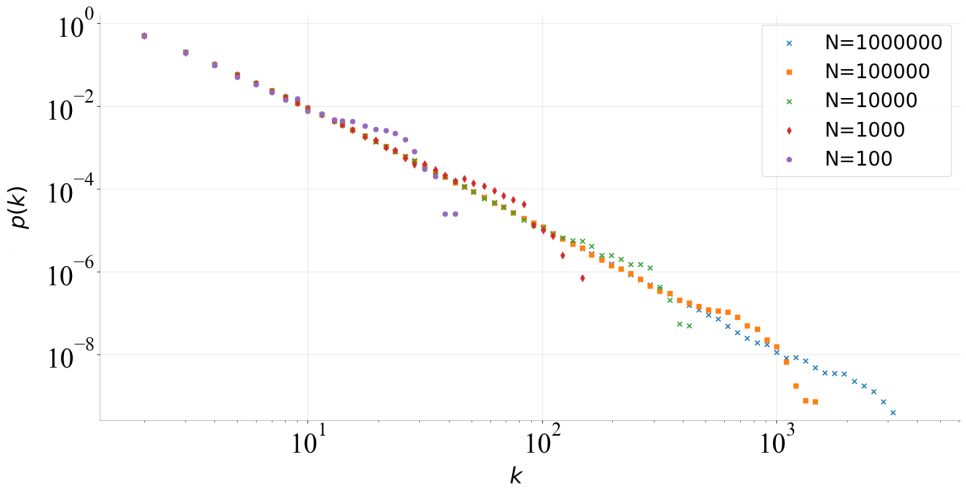


Figure 10: Preferential attachment: log-log plot of $p(k)$ vs. k for different Ns. The data collapse of Figure 11 is performed on these data.

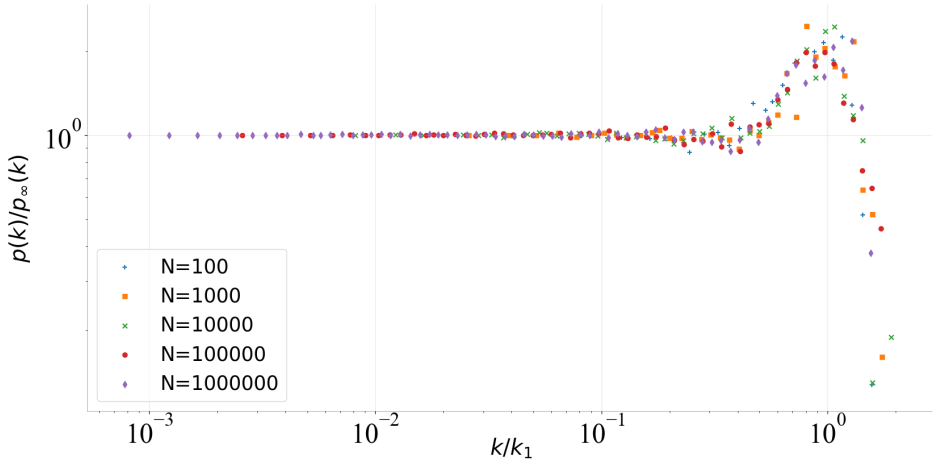


Figure 11: Preferential attachment data collapse: log-log plot of $\frac{p(k)}{p_\infty(k)}$ vs. $\frac{k}{k_1}$ for $N = 10^2 \dots 10^6$.

3 Phase 2: Pure Random Attachment Π_{rnd}

3.1 Random Attachment Theoretical Derivations

3.1.1 Degree Distribution Theory

Using the notation of section 1.1, in the Random Attachment (RA) model, the transition probability is:

$$\mathbb{P}((t+1,s) \in G_{t+1} | G_t) := \Pi_{\text{rnd}}(k,t) := \frac{1}{N(t)}. \tag{30}$$

Let's consider the conditional expected value of $n(k,t)$:

$$\begin{aligned} \mathbb{E}(n(k,t+1)|G_t) &= n(k,t) + m\Pi_{\text{rnd}} n(k-1,t) - m\Pi_{\text{rnd}} n(k,t) + \delta_{k,m} \\ &= n(k,t) + \frac{mn(k-1,t)}{N(t)} - \frac{mn(k,t)}{N(t)} + \delta_{k,m} \end{aligned} \tag{31}$$

We can express it in terms of the probability distribution $p(k,t)$ and look for solutions that are stable in the long-time limit:

$$\begin{aligned}(N(t)+1)\mathbb{E}(p(k,t+1)|G_t) = N(t)p(k,t) + mp(k-1,t) - mp(k,t) = \delta_{k,m} \overset{t\to\infty}{=} \\ \overset{t\to\infty}{=} N(t)p_\infty(k) + \frac{1}{2}(k-1)p_\infty - \frac{1}{2}kp_\infty(k) + \delta_{k,m}\end{aligned} \tag{32}$$

If the stationary solution exists then it is completely determined by the initial conditions and we have

$$\mathbb{E}(p_\infty(k)|G_t) = p_\infty(k). \tag{33}$$

The master equation for the (RA) model is therefore

$$p_\infty(k) = p_\infty(k) + \frac{1}{2}(k-1)p_\infty(k-1) - \frac{1}{2}kp_\infty(k) + \delta_{k,m} \tag{34}$$

Using $p_\infty(k<m) = 0$ we find the recursion formula

$$\begin{aligned}p_\infty(k) &= \left(\frac{m}{m+1}\right) p_\infty(k-1), \text{ for } k > m \\ p_\infty(m) &= \frac{1}{m+1}\end{aligned} \tag{35}$$

so that the exact degree distribution $p_\infty^{rnd}(k)$ in the long-time limit for the random attachment model is

$$p_\infty^{\text{rnd}}(k) = \left(\frac{m}{m+1}\right)^{k-m}\left(\frac{1}{m+1}\right), \text{ for } k \geq m. \tag{36}$$

and it is normalized:

$$\begin{aligned}&\sum_{k=m}^{\infty} n(k) = \sum_{k=m}^{\infty}\left(\frac{m}{m+1}\right)^{k-m}\left(\frac{1}{m+1}\right) = \\ &= \frac{1}{m+1}\sum_{k=m}^{\infty}\left(\frac{m}{m+1}\right)^{k-m} = \\ &= \frac{1}{m+1}\sum_{k'=0}^{\infty}\left(\frac{m}{m+1}\right)^{k'} = \\ &= \frac{1}{m+1}\left(\frac{1}{1-\frac{m}{m+1}}\right) = 1.\end{aligned} \tag{37}$$

3.1.2 Largest Degree Theory

We can set up the same differential equation of Eq. (23):

$$\frac{dk_s}{dt} = m\Pi_{\text{rnd}}(k_s,t) = \frac{m}{2mt-m} \overset{t>>1}{=} \frac{1}{2t}. \tag{38}$$

The solution of this differential equation is:

$$k_s(t) = \frac{1}{2}\ln\left(\frac{t}{t_0}\right) + m. \tag{39}$$

so that $k_s(0) = m = m_0$. We expect the largest degree in the RA model to grow as

$$k_1 \propto \ln N \tag{40}$$

By looking at the expression of the rank of a vertex k we find

$$\begin{aligned} r(k) &= \sum_{k'=k}^{\infty} n(k') = \sum_{k'=k}^{\infty} N p_{\infty}^{\mathrm{rnd}}(k') = \\ &= \left(\frac{N}{1+m}\right) \sum_{k'=k}^{\infty} \left(\frac{m}{m+1}\right)^{k'-m} = \\ &= \left(\frac{N}{1+m}\right) \left(\frac{m}{1+m}\right)^{k-m} \sum_{i=0}^{\infty} \left(\frac{m}{m+1}\right)^{i} = \\ &= \left(\frac{N}{1+m}\right) \left(\frac{m}{1+m}\right)^{k-m} (m+1)\,. \end{aligned} \tag{41}$$

By looking for the degree of the vertex with rank 1 we find:

$$\begin{aligned} \left(\frac{m}{m+1}\right)^{k_1} &= \frac{1}{N} \left(\frac{m}{m+1}\right)^{m} \\ k_1 &= \frac{\ln(N)}{\ln\left(\frac{m+1}{m}\right)} + m. \end{aligned} \tag{42}$$

The Largest degree in a finite system depends on N in the form of Eq. (40). The largest degree is:

$$k_1(N) = \frac{\ln(N)}{\ln\left(\frac{m+1}{m}\right)} + m \tag{43}$$

3.2 Random Attachment Numerical Results

3.2.1 Degree Distribution Numerical Results

The same statistical analysis of Section 2.3.3 has been performed on the degree distribution of the random attachment model. Figure 12 and 13 show the probability distributions and the frequencies distributions (with $n(k) > 5$) used for the statistical analysis. The results of the χ^2-test and the KS test are reported in Table 5.

3.2.2 Largest Degree Numerical Results

The study of the behaviour of k_1 and N consists of a linear regression of the set of points $\{(\ln N, k_1)\,,$ for N $= 10^2, 10^3, 10^4, 10^5, 10^6\}$. Each degree is the result of the average over $M = 100$ realizations, so the errors have been estimated as $\frac{\sigma}{\sqrt{M}}$, with σ being the standard deviation. Figure 14 shows the data with the error bars, together with the regression line and the theoretical behaviour of Eq. (43). Table 6 summarizes the result of the fit.

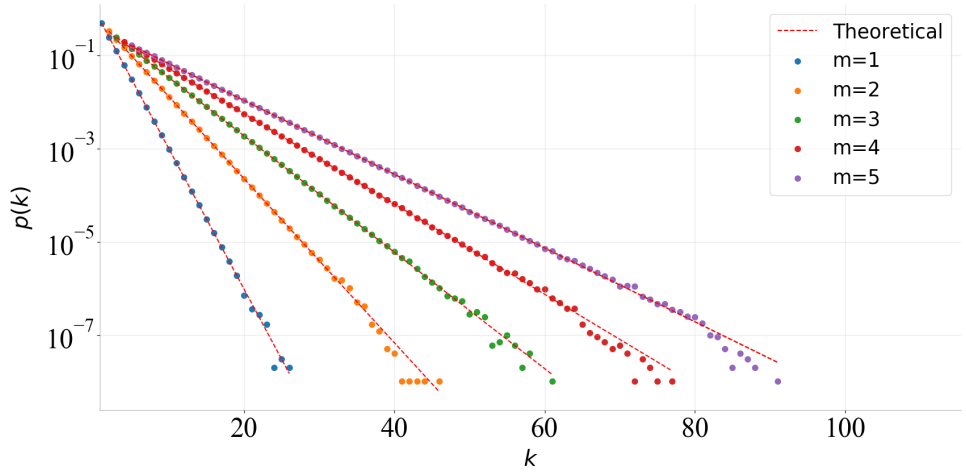


Figure 12: Random attachment: numerical result, in a linear-log plot, of the degree distribution for fixed $N = 10^{10}$ and $m = 1, 2, 3, 4, 10$, compared with the theoretical results. The numerical data is the result of $M = 100$ realizations of the system.

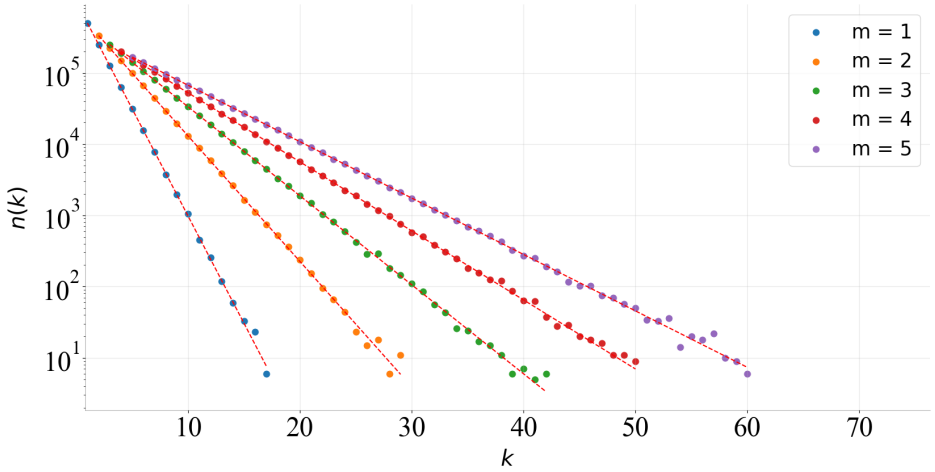


Figure 13: Random attachment: log-log plot of the frequency distributions used for the χ^2-test. The frequency needs to be larger than 5.

m	χ^2	p_{χ^2}	D	p_D
1	17.6	0.35	0.0005	0.999
2	13.8	0.61	0.0004	1.000
3	24.9	0.95	0.0005	1.000
4	403	0.67	0.0004	1.000
5	52.7	0.49	0.0007	0.980

Table 5: Results of the statistical tests for random attachment. The null hypothesis are that the numerical distributions can be described by the theoretical ones (for χ^2-test) and that the two samples come from the same population (KS test). For both we can accept H_0 with a high significance level.

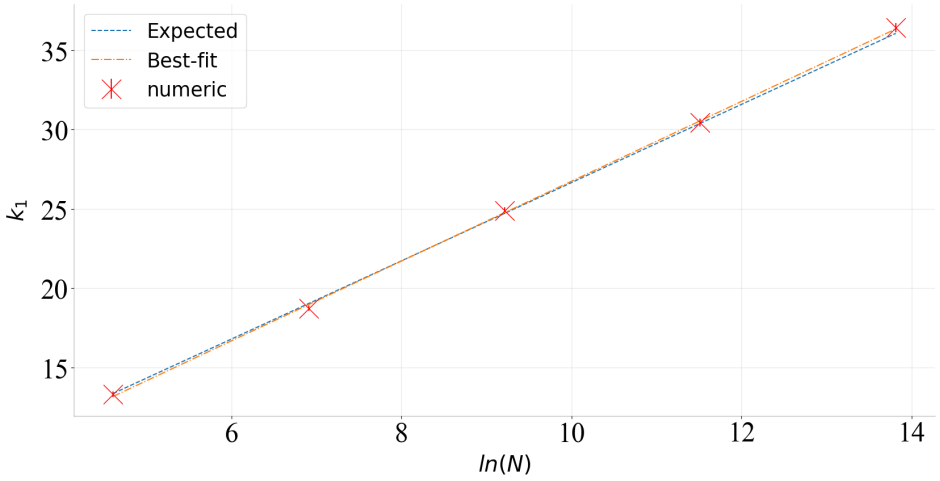


Figure 14: Random attachment: linear-log plot of the largest degree data with the error bars, together with the regression line and the expected behavior: $k_1 = 0.25 \log_{10} N + 2$.

σ_1	σ_2	σ_3	σ_4	σ_5	χ^2	p_{χ^2}	c_1	c_0
0.18	0.16	0.21	0.24	0.29	3.37	0.50	2.52 ± 0.03	1.58 ± 0.3

Table 6: Errors and results of the χ^2-test for the largest degree in random attachment. The last two columns are the parameters of the regression line $c_1 \ln(N) + c_0$. The expected values are $c_1 = 0.25$ and $c_0 = 2 = m$.

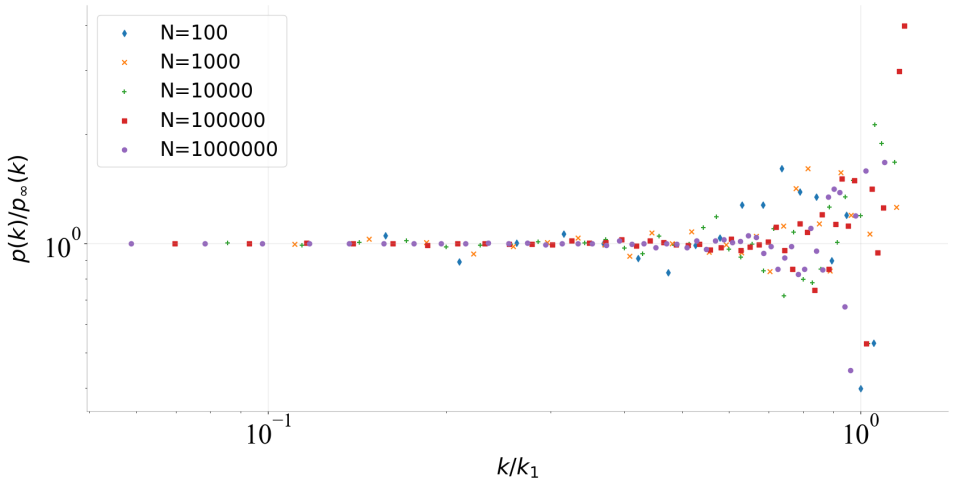


Figure 15: Random attachment data collapse: log-log plot of $\frac{p(k)}{p_\infty(k)}$ vs. $\frac{k}{k_1}$ for $N = 10^2 \ldots 10^6$.

4 Phase 3: Random Walks and Preferential Attachment

4.1 Implementation

In the programme, the graph is defined as described in section 2.1.1. As in the previous models, we add one new vertex each time step, and m edges are connected from this new vertex to the existing ones. The target vertices are chosen using a random walk. The algorithm for the random walk is as follows. The starting vertex is chosen randomly as $i \in \{0, 1, 2, \ldots, |Nb| - 1\}$. At this point, the walk terminates with probability $1 - q$, $q \in [0, 1]$. If the walk does not terminate, a vertex is randomly chosen from the list $Nb[i]$ and the process starts again. The target vertex for the new edge is the last vertex visited in the random walk. An indefinite number of random walks are performed until there are m different targets for the m edges to be added.

4.2 Numerical results

Figure 16 shows a log-log plot of the degree distribution for $N = 10^5$ and different values of q together with the pure random attachment and pure preferential attachment theoretical distributions. A two-sample KS test has been performed for values of q close to 1 under the same hypothesis of section 2.3.3. The results have been reported in Table 7.

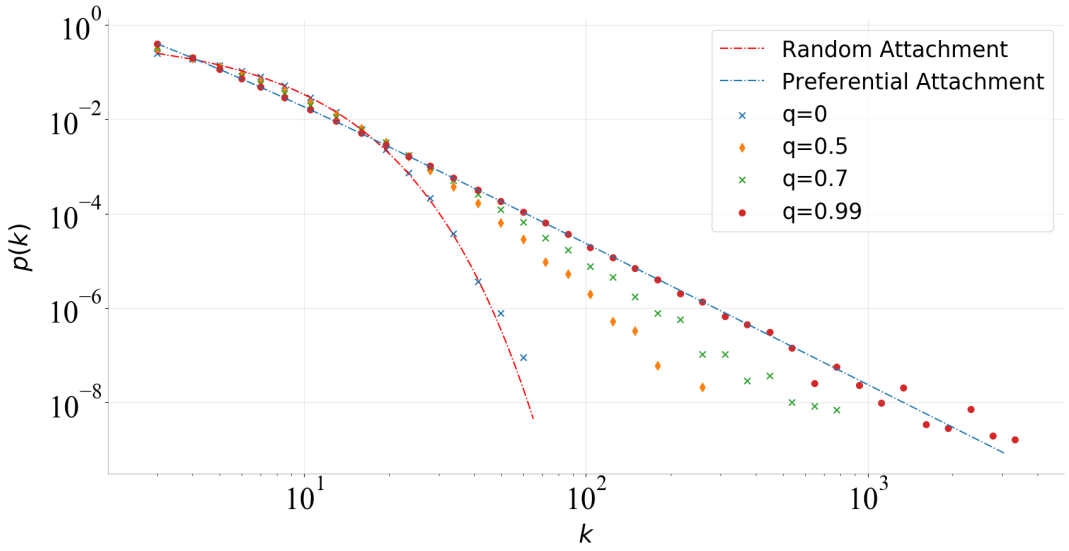


Figure 16: Random walk: log-log plot of the degree distributions for different values of q. The dashed lines are the theoretical distributions of the random attachment and the preferential attachment.

4.3 Discussion of Results

The results can be interpreted in terms of Markov processes. Let $(X_t)_{t=0,\ldots,\infty}$ be a Markov chain, namely a descrete-time stochastic process that satisfies:

$$\mathbb{P}(X_{t+1} = j | X_t = i_t, X_{t-1} = i_{t-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{t+1} = j | X_t = i). \tag{44}$$

q	D	p_D
0.800	0.038	0.0
0.970	0.004	0.33
0.990	0.002	0.99

Table 7: Results of the two-sample KS test. For values of q larger than 0.97 we can accept, with a high significance level, the hypothesis that the two samples come from the same population.

A Markov chain is completely determined by the transition matrix:

Definition 5. *The **transition matrix** of a Markov chain is:*

$$T_{i,j} := \mathbb{P}(X_{t+1} = i | X_t = j) \geq 0.$$

Given a network with adjacency matrix A, a uniform probability distribution on $Nb[i]$ for all i (i.e. the set of vertices linked to i, making use of the notation from section 4.1) defines a Markov chain with transition matrix

$$T_{i,j} = \frac{A_{i,j}}{k_j}, \tag{45}$$

that defines a random walker on the network itself. If we consider the probability distribution of the random variable X_k:

$$\begin{aligned} \boldsymbol{\pi}^{(k)} &= (\pi_i^{(k)}, \ldots, \pi_n^{(k)}) \\ \pi_i^{(k)} &= \mathbb{P}(X_k = i), \end{aligned} \tag{46}$$

the following property holds:

Property 1. *For $\boldsymbol{\pi}^{(t)}$, probability distribution on the set of values of X_t:*

$$\boldsymbol{\pi}^{(t+1)} = T\boldsymbol{\pi}^{(t)} \Leftrightarrow \boldsymbol{\pi}^{(t)} = (T)^t \boldsymbol{\pi}^{(0)}$$

The following theorem holds:

Theorem 1. *Any (regular) Markov chain has a unique probability density vector $\boldsymbol{\pi}$ such that:*

- $\boldsymbol{\pi} = T\boldsymbol{\pi}$
- $\lim_{t\to\infty} \boldsymbol{\pi}^{(t)} = \lim_{t\to\infty} (T)^t \boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}\ \forall \boldsymbol{\pi}^{(0)}$.

The results of the numerical simulation of this model can be interpreted as follow. When $1-q$, namely the probability that the walk ends, is close to zero, the random walk defined by the transition matrix of Eq. (45) lasts for a large number of steps, and $\boldsymbol{\pi}^{(t)}$ get closer to the stationary probability density. It turns out that

$$\lim_{t\to\infty} \boldsymbol{\pi}^{(t)} = \boldsymbol{\pi} = (\frac{k_0}{2E}, \frac{k_1}{2E}, \ldots, \frac{k_n}{2E}) \tag{47}$$

where $E = \sum_s k_s$ is the total number of edges. This is exactly the probability distribution of the BA model ($\Pi_{\text{pa}}(k_s, t)$). For this reason, when q is close to 1, the degree probability

density $p(k)$ follows the same power-law of the BA model. This is related to what the PageRank algorithm does in order to rank the web pages in Google search engine results. Roughly speaking, the pages are given a number that is related to the probability that a random walker will end its (long-time) walk on that page. The walk is performed on a subgraph of the WWW that is the subgraph of web pages that contain the keywords typed in the search engine. The idea is that the higher the number of links to a page, the higher is the importance the of page. This is a simplified version of the random walker defined in the PageRank algorithm, for it does not allow for "hyperjumps", and the edges are not directed.

5 Conclusions

Self-organization and scale invariance are common features of a number of complex systems just as growth and preferential attachment are mechanisms proper of many real-world networks. The Barabási and Albert model, with its preferential attachment, explains the presence of nodes with large connectivities and the fat-tail of the degree distribution. With pure random attachment, the decay is exponential. If the target node is chosen by means of a random walk, then the longer the process lasts, the closer the degree distribution gets to the one found for preferential attachment.

References

[1] Barabási AL, Albert R. Emergence of scaling in random networks. science. 1999;286(5439):509–512.