Network dynamics

Complex and Social Networks (CSN)

Clément Herbin, Odysseas Kyparissis

December 10, 2023

Facultat d'Informàtica de Barcelona (FIB)
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Contents

List of Figures

List of Tables

1 Introduction

Network growth models play a pivotal role in understanding the complex dynamics of evolving systems, particularly in the field of network science. In this laboratory session, our focus is on simulating and analyzing the Barabási-Albert (BA) model and its modified versions, with a keen emphasis on unraveling the underlying statistical properties. The BA model, rooted in two fundamental principles of vertex growth and preferential attachment, serves as a foundation for our exploration.

The primary objectives of this laboratory session encompass three key facets. Firstly, we aim to deepen our comprehension of the dynamical principles that underlie the Barabási-Albert model, shedding light on the intricate interplay between vertex growth and preferential attachment. Secondly, the session endeavors to hone our simulation skills, enabling us to adeptly model and analyze evolving networks. Lastly, we seek to apply curve-fitting methods introduced in a previous lab session, thereby engaging in model selection to discern the most fitting representation of the evolving networks.

Noteworthy parameters characterizing the Barabási-Albert model include n_0 denoting the initial number of vertices, and m_0 representing the initial number of edges for each new vertex. Additionally, the initial set of edges (s_0) can significantly influence the model's dynamics [1].

The Barabási-Albert model's distinctive features, such as preferential attachment following a power-law distribution $ki(t) \sim m_0\sqrt{t/t_i}$ and vertex growth exhibiting an inverse cubic relationship ($p(k) \sim k^{-3}$), set the stage for our exploration. However, this session extends beyond the conventional BA model ($BA-PA$), introducing two modified variants — one where preferential attachment is replaced by random attachment ($BA-RA$) and another where vertex growth is suppressed ($BA-NG$).

Anticipated results include a transition from a power-law vertex growth to an exponential distribution in node degree, accompanied by a shift from the square root of t_i to a logarithmic dependence on t_i. It is crucial to exercise caution when examining the evolving network, as the final outcome converges to a complete graph, resulting in a binary distribution where maximum degree yields a probability of 1, and all other degrees yield a probability of 0. From a technical standpoint, generating an average curve for vertices arriving at the same time ensures a smoother representation of the fitted model, mitigating the risk of observing a step function when plotting k_i versus time.

In summary, this laboratory session delves into the exploration of network growth models, particularly the Barabási-Albert model and its modifications. Through rigorous simulation and analysis, we endeavor to unravel the intricate statistical properties governing evolving networks, providing insights into the dynamical principles that shape their structure.

2 Results

In this section, the obtained results are presented. We choose to use $m_0 = 2$ and $t_{max} = 10000$. We also choose to use $n_0 = 250$ for BA-PA and BA-RA and $n_0 = 1000$ for BA-NG. Firstly, details of the degree sequence analyses are provided, followed by the results on the vertex growth topic.

2.1 Degree Sequence Results

To begin with, Table1 is introduced to showcase the different basic statistics of each degree sequence for the three variations of the BA model ($BA - PA$, $BA - RA$, $BA - NG$). This table summarizes the degree properties, as well as, some statistical information about the BA networks generated.

Version	**N**	**Max. degree**	**M/N**	**N/M**
BA-PA	10250	65	3.95122	0.2530864
BA-RA	10250	18	3.95122	0.2530864
BA-PA-NG	1000	999	41.91800	0.0238561

Table 1: Summary of the properties of the degree sequences. N is the number of nodes, M is the sum of degree.

As mentioned before, an ensemble of degree distributions is considered, on which model selection is performed to estimate the degree distribution of each version of the BA models that were generated during the simulation phase. In all cases it is defined that $p(0) = 0$. Table2 displays the ensemble of probability distributions used.

Probability distribution	**Model ID**
Displaced Poisson	D1
Displaced geometric	D2
Zeta with $\gamma = 3$	D3
Zeta	D4
Right-trucanted zeta	D5

Table 2: Proposed probability distributions

The ensemble contains two distributions from null models of networks and two nested variants of the zeta distribution as possible models of power-laws. The distributions deriving from null models of networks are the Poisson and the geometric distributions. The Poisson distribution is chosen for being a mathematically simple approximation to the binomial distribution characterizing Erdos-Renyi graphs. The formulas of the distributions can be found in equations 6, 7, 8, and 10.

Before proceeding further, bar plots of the degree distribution per version of the BA model, in log-log scale, can be observed in Figures 1, 2, 3. The scale-free nature of the first two networks is already evident, while for $BA - NG$ the pattern is completely different.

2.1.1 Estimation of the parameters

Before applying standard model selection methods, the parameters giving the best fit must be obtained. To estimate the parameters, we use the minus log-likelihood function and instead of minimizing it, its maximization takes place. The most likely parameters estimated by the ensemble model are displayed in Table 3. From the information of Table1 and the results of Table 3 it can be observed that $q \approx q_0$ and $\lambda \approx \lambda_0$, while γ and γ_0 in some cases differ substantially.

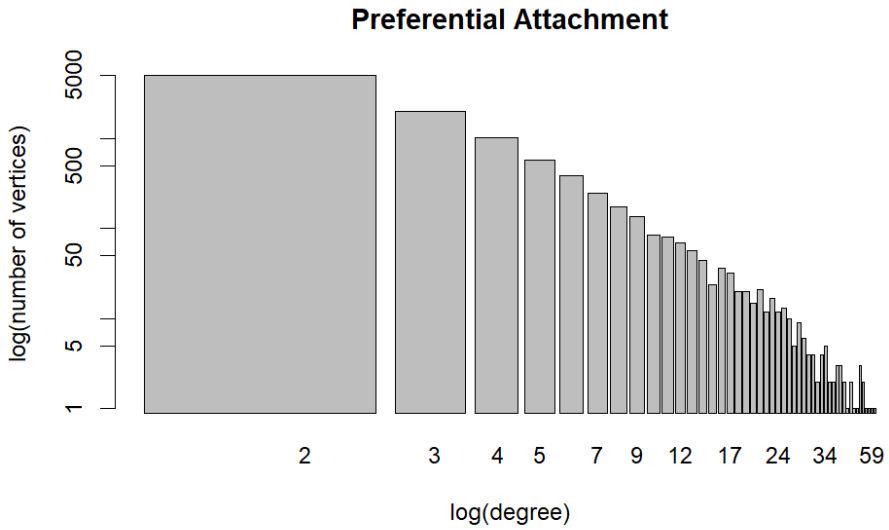

Figure 1: Visualization of the degree distribution of $BA - PA$

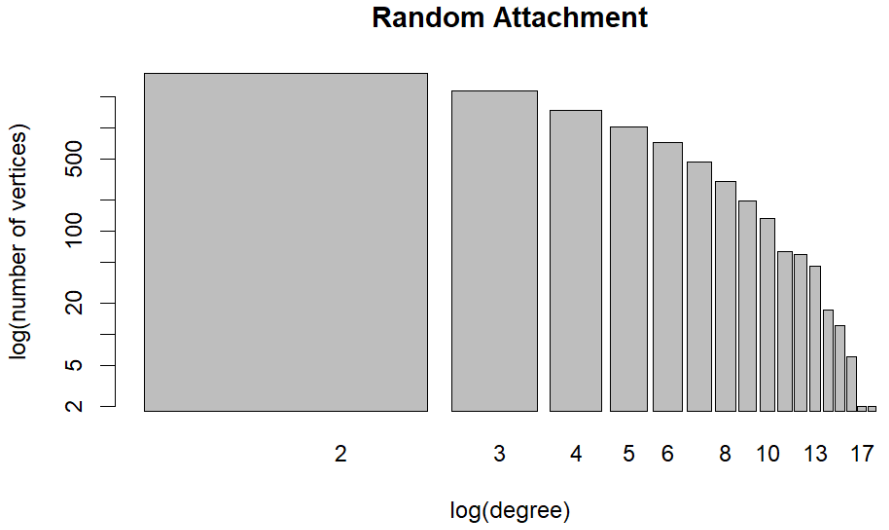

Figure 2: Visualization of the degree distribution of $BA - RA$

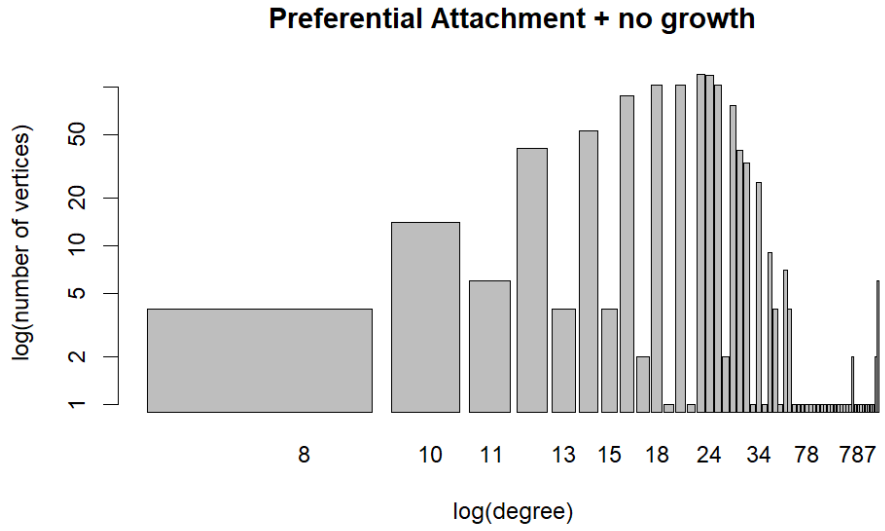

Figure 3: Visualization of the degree distribution of $BA - NG$

Version	**Model D1** (λ)	**Model D2** (q)	**Model D4** (γ_1)	**Model D5** (γ_2)	(k_{max})
BA-PA	3.868696	0.2530864	1.619799	1.413585	65
BA-RA	3.868696	0.2530864	1.580940	1.0	18
BA-PA-NG	41.918000	0.0238561	1.271104	1.004111	999

Table 3: Summary of the most likely parameters, γ_1 and γ_2 refer respectively, to the exponent of the zeta distribution and the right-truncated distribution. See Table 2 for the key of models ID.

Each of the probability distributions included in the ensemble has been a candidate to fit the degree distribution of the BA simulations. The best one has been chosen according to the Akaike information criterion (AIC); more in section 5. The AIC difference of all tested models is displayed in Table 4.

Model Version	**Model D1**	**Model D2**	**Model D3**	**Model D4**	**Model D5**
BA-PA	42898.41	96152.08	0	104513.7	102789.8
BA-RA	49205.20	100356.61	0	112081.7	105568.9
BA-PA-NG	0.00	232481.36	210031	233974.5	233377.0

Table 4: AIC difference of a model for the degree sequence of the different version of the BA model. See Table 2 for the key of models ID.

The best fitting model per version, is shown in Figures 4, 5 and 6. Those figures display the results obtained with the ensemble model estimation for all the available versions.

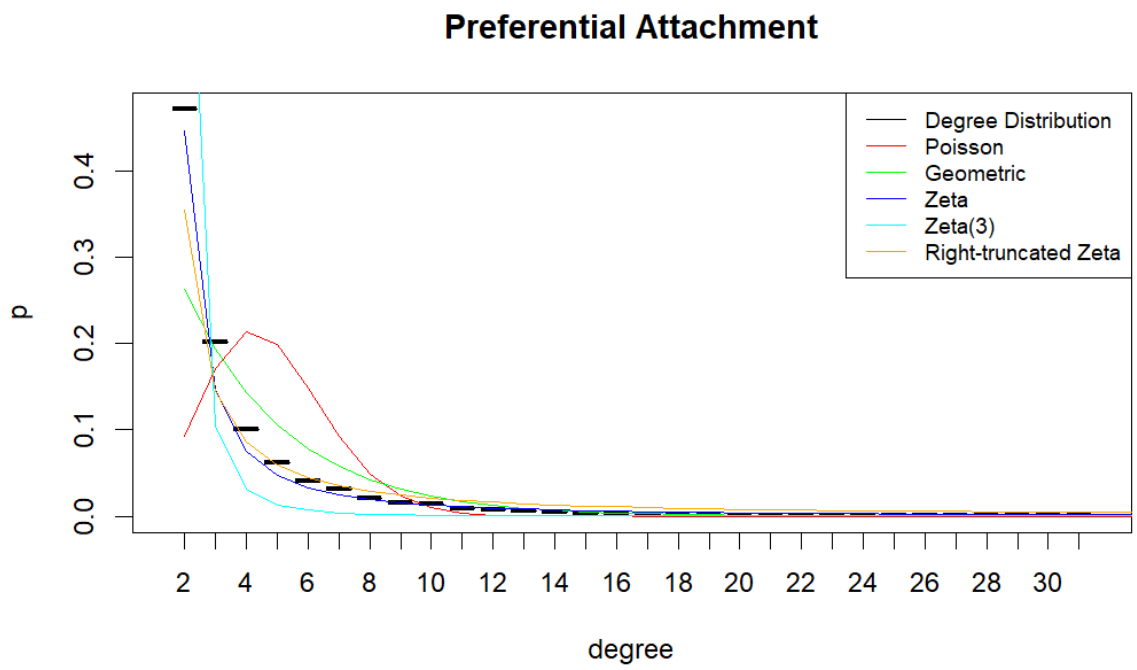

Figure 4: Results for $BA - PA$ degree distribution

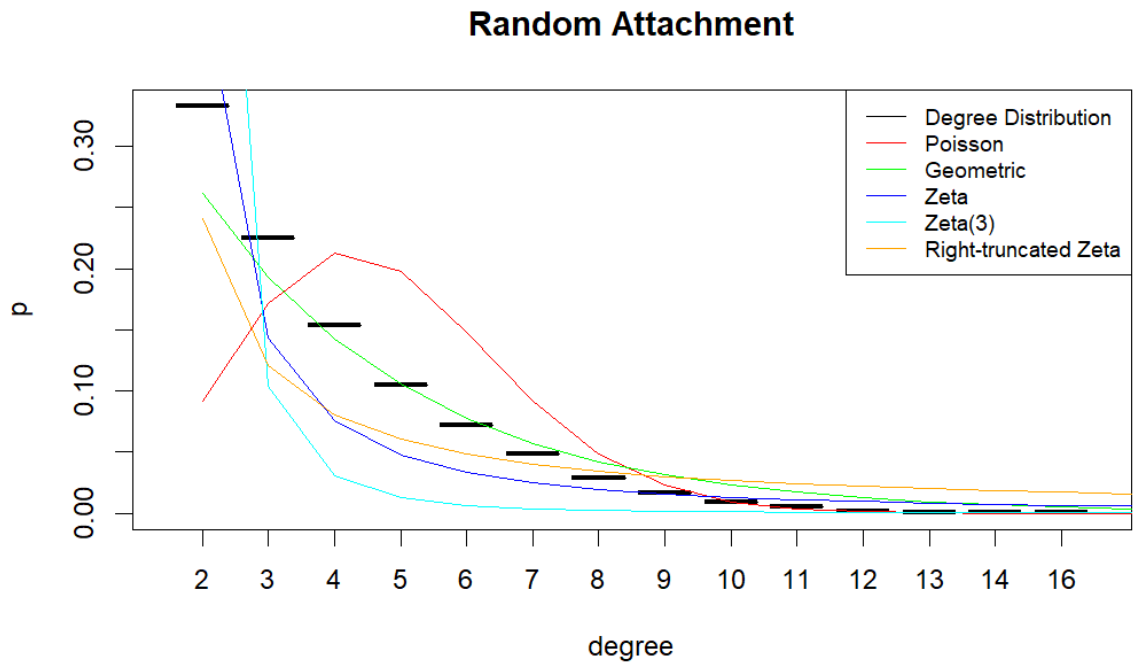

Figure 5: Results for $BA - RA$ degree distribution

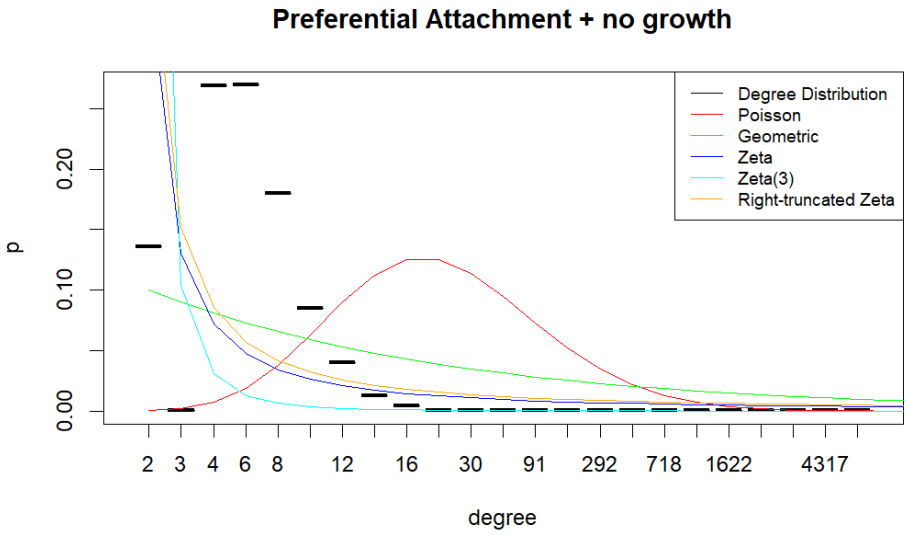

Figure 6: Results for $BA - NG$ degree distribution

2.2 Vertex Growth Results

In this subsection, firstly, we present visualizations of $k_i(t)$ for all model versions, followed by tables of fitting metrics and parameter estimations, and lastly the final best models are included. For each different model version, 4 distinct vertices were selected, in order to analyse the growth of their degree over time. Specifically, the vertices selected in all cases are the ones that added to the network at the following timesteps: $t_1 = 1$, $t_{10} = 10$, $t_{100} = 100$, $t_{1000} = 1000$. To begin with, information related to the network of each BA model version is presented. The following visualizations (Figures 7, 8, 9) include the plotting of the target metric ($k_i(t)$) with respect to time (t), together with the theoretical distributions described below.

Firstly, concerning the original version of the BA model ($BA - PA$), the growth of k_i the degree of the $i - th$ vertex as a function of time obeys:

$$k_i(t)_{BA-PA} \approx m_0 \left(\frac{t}{t_i}\right)^{\frac{1}{2}} \tag{1}$$

although a re-scaled variant calculated as:

$$k'_i(t)_{BA-PA} = \frac{t^{1/2}}{t_i} k_i(t) \approx m_0 t^{1/2} \tag{2}$$

should be about the same for every vertex, regardless of its arrival time. Thus, this is presented in Figure 7. As for the random-attachment version of the BA model ($BA - RA$), the growth of k_i the degree of the $i - th$ vertex as a function of time obeys:

$$k_i(t)_{BA-RA} \approx m_0 \left(\log(m_0 + t - 1) - \log(n_0 + t_i - 1) + 1\right) \tag{3}$$

where a re-scaled variant calculated as:

$$k''_i(t)_{BA-RA} = k_i(t) + m_0 \log(n_0 + t_i - 1) - m_0 \approx m_0 \log(m_0 + t - 1) \tag{4}$$

should be again about the same for every vertex, regardless of its arrival time. This result is presented in Figure 8. Finally, if growth is removed but preferential attachment is retained

($BA-NG$ version), the growth of k_i the degree of the $i-th$ vertex as a function of time obeys (for large $n0$ and large t; $t \geq n0$ is required):

$$k_i(t)_{BA-NG} \approx \frac{2m_0}{n_0}t \tag{5}$$

In this case, as all vertices arrive at the same time, Eq. 5 indicates that all vertices grow approximately in the same fashion (on average) for sufficiently large t. This result is presented in Figure 9.

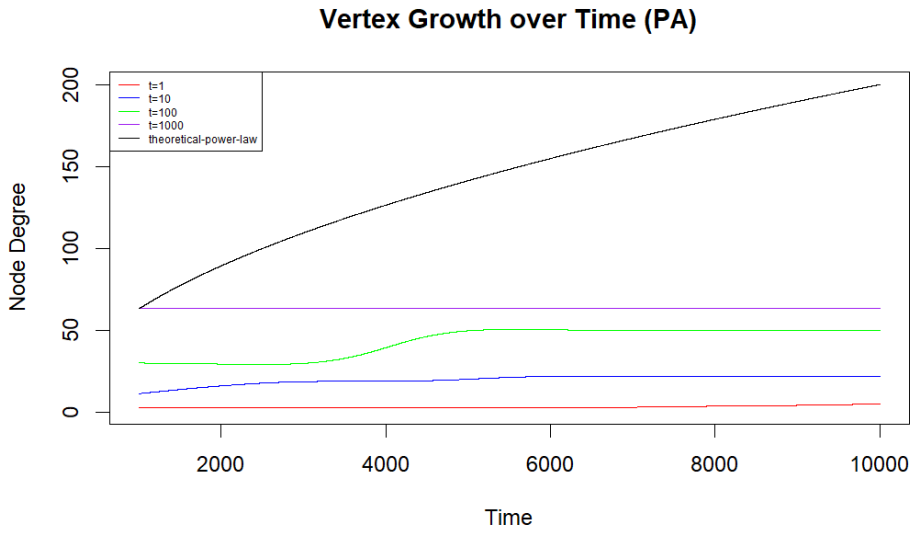

Figure 7: Vertex growth of $BA-PA$ nodes

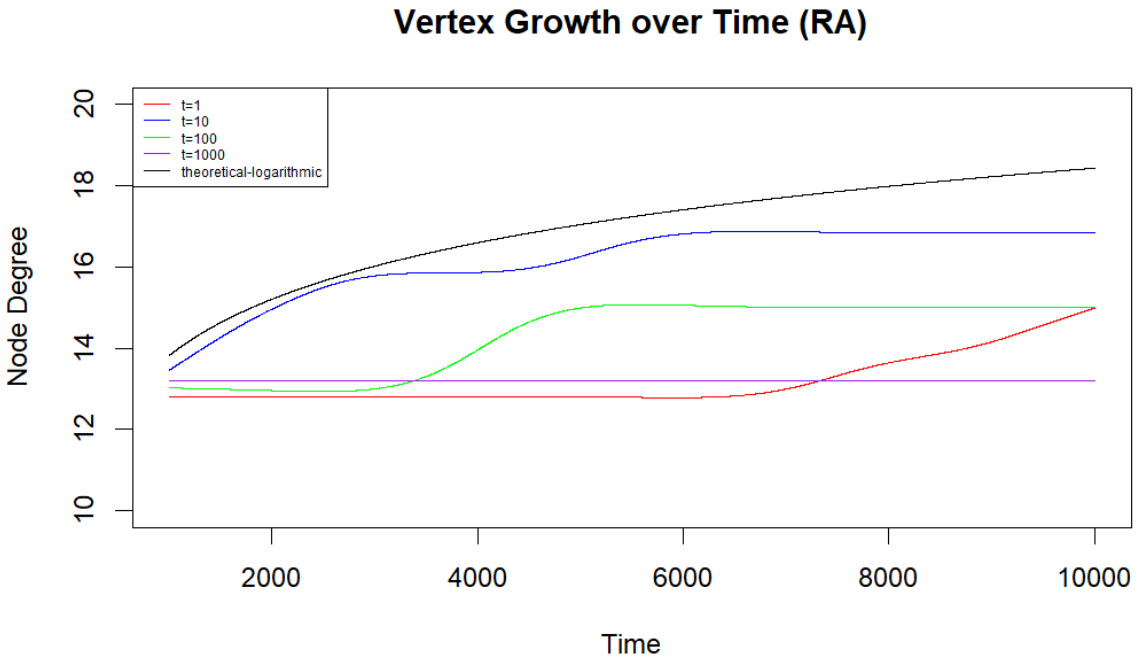

Figure 8: Vertex growth of $BA-RA$ nodes

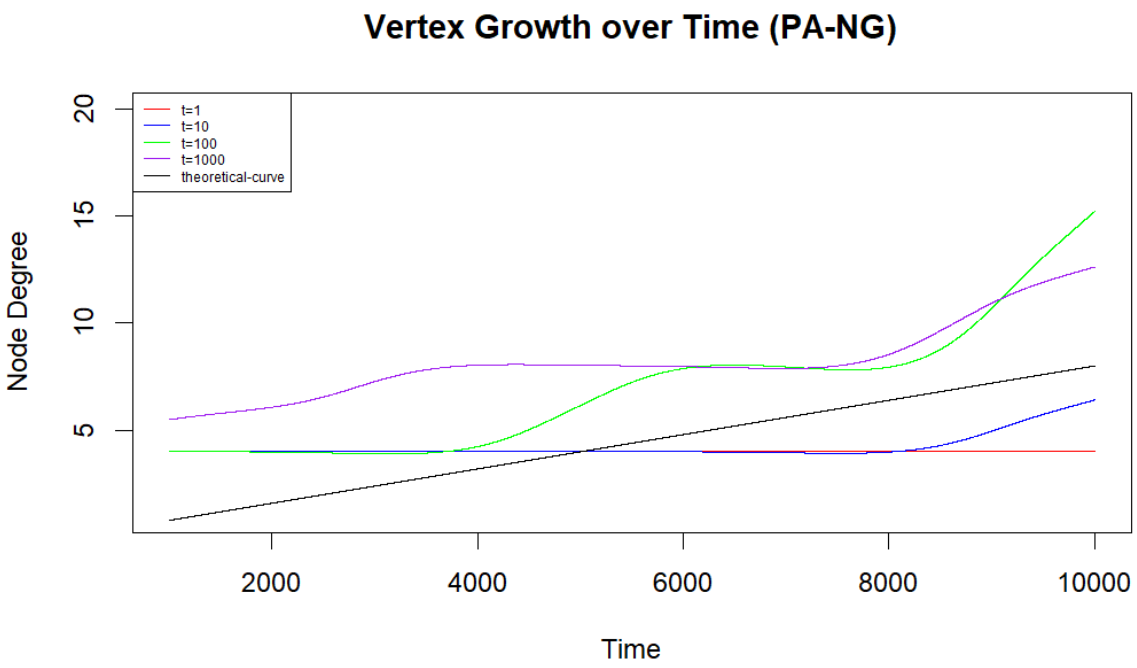

Figure 9: Vertex growth of $BA - NG$ nodes

2.2.1 Model fitting

Before getting in greater details on the results of model fitting, it is crucial to mention that the goal of this section was to find the best fit of the time-series data of the degree growth for the specific 4 vertices given time t from an ensemble of models. The models used are described in Section 5, more specifically under equations: 14, 15, 16, 17, 18, 19, 20, 21, 22. However, in this specific implementation, the models 19 and 20 were facing technical difficulties during fit, and they are considered as future work.

2.2.2 Model parameter estimation

In this subsection, the final parameters for each model for all versions of the BA model are included. Tables 5, 6, 7 and 8 presents the estimated parameters.

Model	0	1	2		4		0+		1+		2+			4+		
	a	a	a	b	a	d1	a	d	a	d	a	b	d	a	d1	d2
BA-PA	0.0017	0.14	0.46	0.36	1.25	0	0.0007	6.2	0.11	2.46	63.4	0.04	-79	1.25	0	0
BA-RA	0.0018	0.15	1.6	0.23	1.36	0	0.0005	8.5	0.078	5.86	47.9	0.037	-54	1.3	0	0
BA-PA-NG	0.0029	0.24	0.012	0.83	2.02	0	0.002	2.2	0.36	-9.1	0.0008	0.87	0.63	2.02	0	0

Table 5: Parameter values for all models by BA model version with t=1

Model	0	1	2		4		0+		1+		2+			4+		
	a	a	a	b	a	d1	a	d	a	d	a	b	d	a	d1	d2
BA-PA	0.001	0.09	0.23	0.39	0.78	0	0.0005	3.6	0.073	1.29	0.15	0.42	0.52	0.78	0	0
BA-RA	0.001	0.09	3.09	0.101	0.85	859	0.0001	6.3	0.023	5.6	2.7e-07	1.68	6.7	0.85	859	0
BA-PA-NG	0.003	0.245	0.018	0.79	2.05	0	0.002	2.5	0.36	-9.1	0.4	0.48	-9.6	2.0	0	0

Table 6: Parameter values for all models by BA model version with t=10

Model	0	1	2		4		0+		1+		2+			4+		
	a	a	a	b	a	d1	a	d	a	d	a	b	d	a	d1	d2
BA-PA	0.0007	0.06	0.01	0.69	0.51	0	0.0005	1.3	0.07	-0.74	4e-10	2.52	2.68	0.5	0	0
BA-RA	0.001	0.089	2.85	0.10	0.8	0	0.0001	6.1	0.19	5.38	8.81	0.47	-6.4	0.71	0	0
BA-PA-NG	0.002	0.22	0.07	0.62	1.8	0	0.0019	4.88	0.27	4.88	1275	0.006	-1334	1.8	0	0

Table 7: Parameter values for all models by BA model version with t=100

Model	0	1	2		4		0+		1+		2+			4+		
	a	a	a	b	a	d1	a	d	a	d	a	b	d	a	d1	d2
BA-PA	0.0005	0.05	0.83	0.17	0.44	0	0.0001	3.03	0.019	3.03	3.9	0.06	-3.55	0.44	0	0
BA-RA	0.001	0.089	2.85	0.10	0.8	0	0.0001	6.1	0.19	5.38	8.81	0.47	-6.4	0.71	0	0
BA-PA-NG	0.0008	0.071	0.54	0.26	0.62	0	0.0002	3.65	0.04	2.31	9.89	0.07	-13.1	0.62	0	0

Table 8: Parameter values for all models by BA model version with t=1000

2.2.3 Model metrics

Moreover, Tables 9, 10, 11 and 12, present the *AIC* values achieved for each distinct model, for all different versions of the BA model. It can be seen that the ***best*** model (the one with lowest *AIC*) is **Model** 2+ for all versions except for the model BA-RA at $t = 100$, which the best model is **Model 4+** and the BA-PA at $t = 1000$ which the best model is **Model 2**.

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	44515	26789	21993	34101	26526	23019	21386	34103
BA-RA	49050	33728	10774	21965	18055	13654	10163	21967
BA-PA-NG	32788	43180	27430	57233	27714	30211	27400	57235

Table 9: AIC metrics for all models with t = 1

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	34152	11191	4223	25199	7643	4275	4204	25201
BA-RA	43740	35222	5417	5281	3163	4252	2617	5283
BA-PA-NG	36122	42463	29117	57163	31874	26490	26388	57165

Table 10: AIC metrics for all models with t = 10

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	23119	22251	20702	30677	18213	21777	11296	30697
BA-RA	43161	32223	10971	10694	12509	12295	10924	10570
BA-PA-NG	45739	39458	37733	52739	40159	36812	36522	52741

Table 11: AIC metrics for all models with t = 100

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	23119	22251	20702	30677	18213	21777	11296	30697
BA-RA	43161	32223	10971	10694	12509	12295	10924	10570
BA-PA-NG	45739	39458	37733	52739	40159	36812	36522	52741

Table 12: AIC metrics for all models with t = 1000

Finally, Tables 13, 15, 15 and 16, presents the difference of *AIC* values (Eq. 13) achieved for each distinct model when it is compared with the best one, for all versions of the BA model. So, the best model is the one with $\Delta = 0$.

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	23128	5402	607	12714	5138	1632	0	12716
BA-RA	38886	23564	610	11801	7891	3490	0	11803
BA-PA-NG	5388	15780	29	29833	313	2811	0	29835

Table 13: AIC difference Δ for all models with t = 1

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	29948	6986	19	20995	3439	71	0	20997
BA-RA	41222	29904	2799	2663	545	1634	0	2665
BA-PA-NG	9633	15974	2628	30675	5385	1	0	30677

Table 14: AIC difference Δ for all models with t = 10

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	11822	10954	9406	19381	6916	10481	0	19383
BA-RA	32590	21652	400	124	2938	1725	353	0
BA-PA-NG	9216	2935	1210	16216	3636	289	0	16218

Table 15: AIC difference Δ for all models with t = 100

Model	0	1	2	4	0+	1+	2+	4+
BA-PA	29020	16413	0	1287	3550	1517	14795	1289
BA-RA	29712	14012	165	9200	2794	754	0	9202
BA-PA-NG	11909	14854	9981	25618	7487	12526	0	25620

Table 16: AIC difference Δ for all models with t = 1000

2.2.4 Best fitted models

In this subsection, the presentation of the best fitted models is taking place. Figures 10, 11 and 12 illustrates vertex growth over time, together with the ***best model fit*** (green line), and the theoretical curve (red line) for all versions of the BA model simulated. For the BA-PA model, the theoretical curve is $k_i(t) = m_0 t^{1/2}$, for BA-RA, it's $k_i(t) = m_0 log(m_0 + t - 1)$ and for the BA-NG model, it's $k_i(t) = 2m_0 t/n_0$

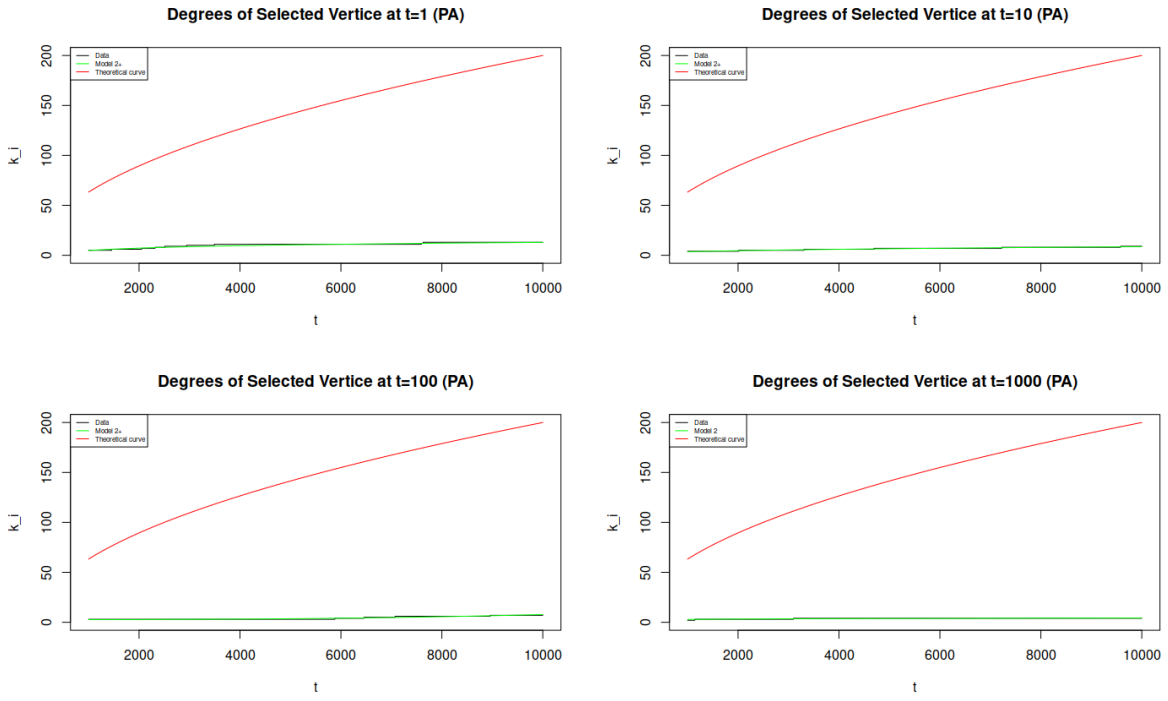

Figure 10: Plots of the data, theoretical curve and best model for the BA-PA model

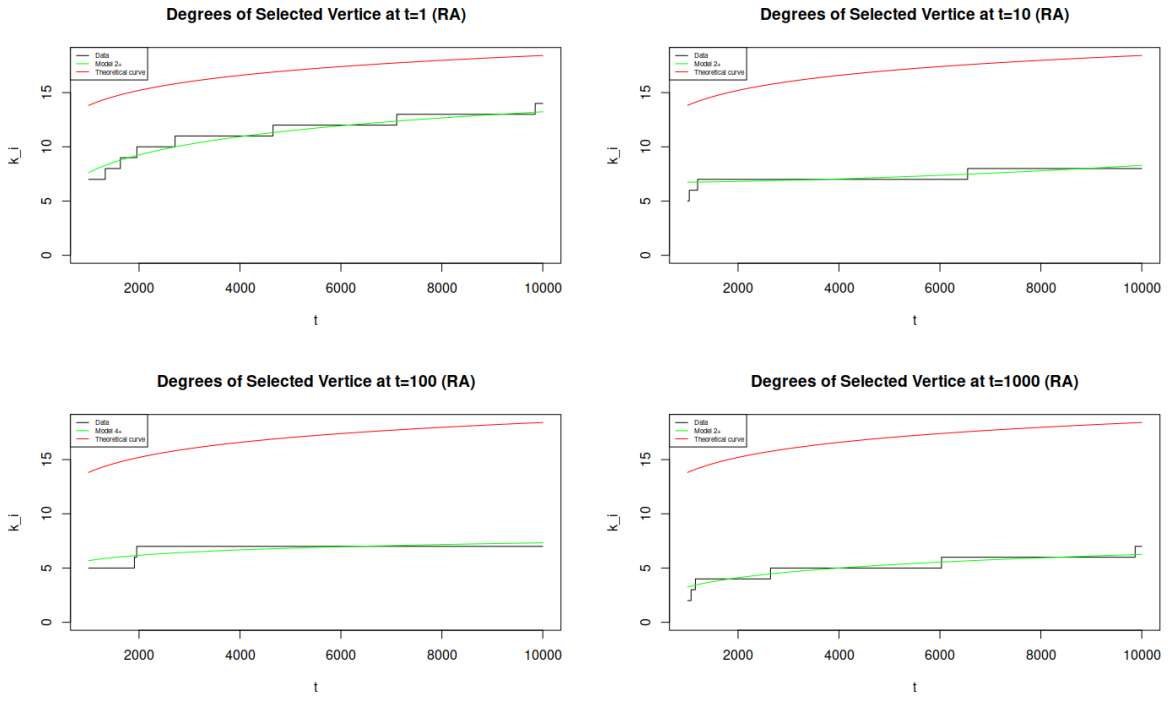

Figure 11: Plots of the data, theoretical curve and best model for the BA-RA model

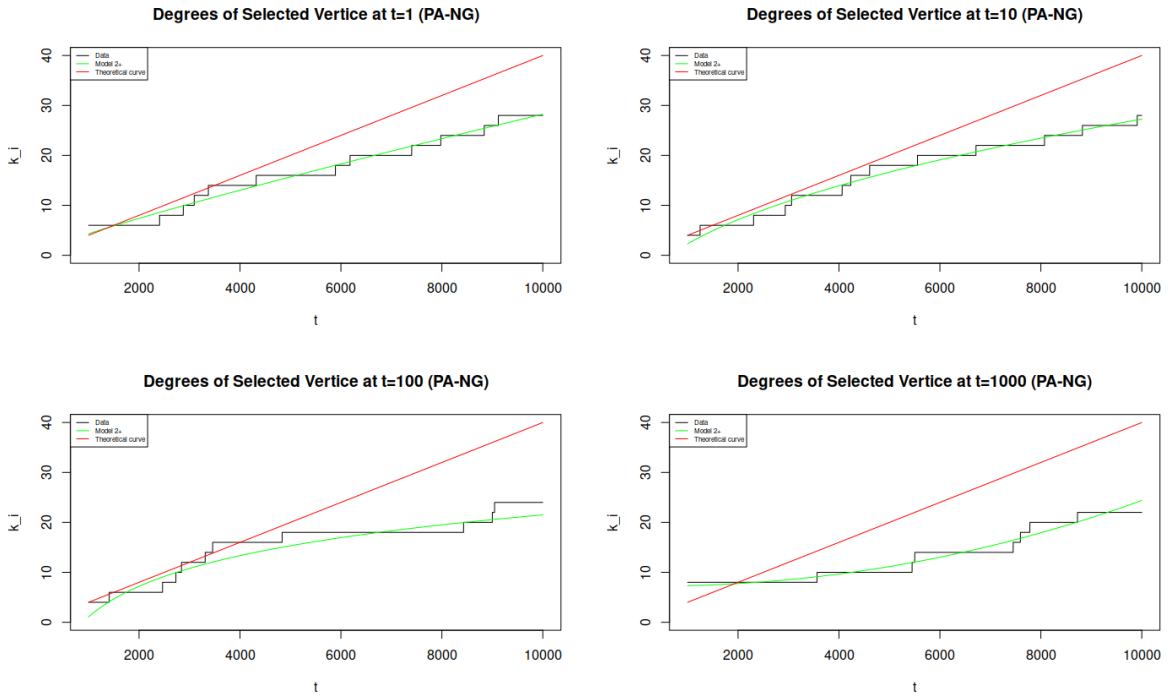

Figure 12: Plots of the data, theoretical curve and best model for the BA-NG model

3 Discussion

3.1 Degree Sequence

The degree sequence analysis provides insights into the structural properties of the Barabási-Albert (BA) model and its variations. The theoretical expectations and the observed results are compared for each version: $BA - PA$, $BA - RA$, and $BA - NG$.

BA-PA: The theoretical expectation for BA-PA suggests a power-law distribution with an exponent of -3, given by $p(k) \approx ck^{-\gamma}$, where $\gamma = 3$ and $c = \frac{2m_0}{n_0+t}$. The expectation is in line with the observed results, as displayed in Table 4 and Figure 4. The theoretical background recommends checking if the best-fitting distribution corresponds to a power-law with a -3 exponent, modeled as a zeta distribution or a right-truncated zeta distribution. The results from Table 4 indicate that the AIC differences for these models are substantial, supporting the choice of the right-truncated zeta distribution.

BA-RA: For this case, it was suggested to check that the best-fitting distribution is no longer a power-law and to compare the quality of the fit with a geometric distribution. The theoretical expectation of a displaced geometric distribution is not consistent with the observed results, as shown in Table 4. However, in Figure 5 it is visually presented that the displaced geometric distribution tends to fit better the data.

BA-NG: The theoretical expectation for BA-NG involves an evolving degree distribution transitioning from a power-law to Gaussian-like and finally to a Kronecker delta function. Although, the expected distribution should be closer to a binomial distribution for sufficiently large t. The results from Table 4 and Figures 6 align with these expectations, indicating that the best-fitting distribution is no longer a power-law for sufficiently large t, but a displaced Poisson.

In summary, the degree sequence analysis, as presented in the results section, demonstrates a medium agreement between the theoretical expectations and the observed results for each version of the BA model. The selected probability distributions provide accurate representations of the degree distributions observed in the simulations.

3.2 Vertex Growth

In this subsection, we delve into a comparative analysis between the obtained vertex growth results and the theoretical expectations for each variant of the Barabási-Albert (BA) model: $BA - PA$, $BA - RA$, and $BA - NG$.

BA-PA: The observed growth patterns in the $BA - PA$ model is not close to the theoretical expectations. The degree $k_i(t)$ demonstrates a growth behavior approximating a power law but with a different exponent from the theoretical power law $k_i'(t) = m_0 t^{1/2}$. The rescaled variant $k_i'(t)$, indicating a common growth trajectory for all vertices, was visually validated in Figure 7. Data are well below expectations. This may be due to the way we simulate the model.

BA-RA: For the BA-RA model, the replacement of preferential attachment with random attachment introduces a logarithmic term in the growth equation. The observed $k_i(t)$ don't aligns well with the theoretical prediction of $m_0 \left(\log(m_0 + t - 1) - \log(n_0 + t_i - 1) + 1\right)$ too. Only 1 of the 4 time series is modeled by a formula including a logarithm (Model 4)+, while the other 3 are power-laws (Model 2+). According to our results, the arrival time is related to the scaling of vertex degree, but theoretically, this is not the case, as demonstrated in Figure 11. This also may be due to the way we simulate the model.

BA-NG: In the BA-NG model, where growth is removed, the observed growth pattern of $k_i(t)$ matches the theoretical expectation better than the other two models, with a theoretical value of $\frac{2m_0}{n_0}t$. Visual checks on the growth trajectory for each vertex indicated a consistent pattern. However, this pattern is not the one we're looking for. The theoretical pattern is a linear model (Model 0 or 0+), which represent the absence of growth, whereas we get another power-law pattern (Model 2+). Theoretical predictions are not aligned with empirical results, as depicted in Figure 12.

Even if for each variant of the BA model, the different time series of the growth of the degrees of a vertex are similar, there is a big difference with the theoretical values.

4 Conclusion

In conclusion, the degree sequence analysis demonstrates a medium agreement between theoretical expectations and observed results for each BA model variant, with selected probability distributions accurately representing degree distributions in simulations. While the time series of growth for each variant of the BA model exhibit similarities, significant differences with theoretical values persist. Further scrutiny of the simulation methodology and model assumptions may be warranted for a more accurate alignment of theoretical predictions with empirical results.

5 Methodology

5.1 BA Simulations

To simulate the Barabási-Albert model and its variations, we employed three separate procedures which respectively were following preferential attachment and time-based growth ($BA-PA$), random attachment and time-based growth ($BA-RA$), and finally preferential attachment with suppressed growth ($BA-NG$). The generation of data involved utilizing a vector of stubs, where an edge connecting vertices u and v contributed two stubs (u and v). At any given time t, prior to introducing a new vertex, the occupied positions in the vector were determined by $s_0 + 2m_0(t-1)$. Here, s_0 denotes the initial number of stubs added at time 0, and $2m_0(t-1)$ represents the cumulative stubs added up to time t. The arrival of a new vertex at time t prompted the random selection of stubs until m_0 edges could be formed. Notably, in models ($BA-PA$ and $BA-NG$) the preferential attachment rule was enforced, ensuring that the probability of selecting a particular vertex was proportional to its degree, although for $BA-PA$ model this procedure was replaced with a random selection of vertex to from the connections. It is worth emphasizing that a vertex's current degree equates to the number of stubs it possesses in the vector. For the initial configuration, each node has a degree equal to $m0$, so we have $s_0 = m_0 n_0$.

5.2 Initial configuration

5.3 Degree Sequence Analysis

To assess the goodness of fit, we compare the observed degree distribution to various probability distributions commonly used in network science, such as the power-law distribution, the zeta family distributions, and others, all listed in Table 2. This comparative approach allows us to identify the most suitable probability distribution for modeling the degree distribution of each BA version among the selected ensemble. This section deep-dives into the details of the procedures carried out to conduct this empirical analysis.

5.3.1 Probability ensemble

The formulas of the different probability distributions tested are presented below. The probabilities expressions for each of the models in the ensemble are detailed:

$$\text{Displaced Poisson} \qquad p(k) = \lambda^k e^{-\lambda}/(k!(1-e^{-\lambda})) \tag{6}$$

$$\text{Displaced geometric} \qquad p(k) = (1-q)^{k-1}q \tag{7}$$

$$\text{Zeta with } \gamma = 3 \qquad p(k) = k^{-3}/\zeta(3) \tag{8}$$

$$\text{Zeta} \qquad p(k) = k^{-\gamma}/\zeta(\gamma) \tag{9}$$

$$\text{Right-truncated zeta} \qquad p(k) = k^{-\gamma}/H(k_{max}, \gamma) \tag{10}$$

As discussed in section 3, the inclusion of several distributions in the ensemble model, allows for a wider scope of tested distributions, and further remarks the notion of the difficulty of finding a perfect-fitting model for the degree distribution of the BA model and its alternative versions.

5.3.2 Estimation of the parameters

By using the log-likelihood formulas for each distinct distribution, certain initial value of the parameter(s), depending on the specific model being tested, and a maximization method, e.g., *LBFGS-B*, the estimation of the optimal distribution and its parameters, per model version, was enabled. The *LBFGS-B* is a minimization/maximization method that allows defining upper

and lower bounds on parameters. Details about the boundaries used during the maximization of the log likelihood process can be observed in the files of the delivered code.

5.3.3 Model selection

To select the parameters we maximize the log-likelihood function (Eq. (11)).

$$\mathcal{L} = \log\left(\Pi_{i=1}^{N} p(k_i)\right) = \Sigma_{i=1}^{N} \log p(k_i). \tag{11}$$

Using the Akaike Information Criterion (AIC) (Eq. (12)), with a correction for the sample size, we obtain the optimal parameters for each tested distribution. To be precise, for the values of N and K that we are going to use ($N \gg K$ in our case), the correction is likely to not alter the conclusions of model selection with regard to the original AIC. AIC is a measure that accounts for the trade-off between minimizing error occurrence and keeping model complexity low.

$$AIC_c = -2\mathcal{L} + 2K\frac{N}{N-K-1}. \tag{12}$$

The AIC difference (Δ) between the lowest AIC achieved per BA model version and the AIC of the remaining distributions included in the ensemble are displayed in Table 4, and it presents the deviations between the different types of distributions on their predictability power for this specific task, also defined in Eq. 13.

$$\Delta = AIC_{\text{model}} - AIC_{\text{best\_model}} \tag{13}$$

5.4 Vertex Growth Analysis

5.5 Ensemble of models

In this section we are going to describe the methodology followed during the realisation of the vertex growth analysis. To begin with, as it was previously mentioned, the goal of the specific task is to find the best fit from an ensemble of models, that will describe the scale of the degree (k_i) of a node v_i over time t. The expressions of the models' functions in the ensemble are detailed:

$$\text{model 0} \qquad f(t) = at \tag{14}$$

$$\text{model 1} \qquad f(t) = at^{1/2} \tag{15}$$

$$\text{model 1+} \qquad f(t) = at^{1/2} + d \tag{16}$$

$$\text{model 2} \qquad f(t) = at^{b} \tag{17}$$

$$\text{model 2+} \qquad f(t) = at^{b} + d \tag{18}$$

$$\text{model 3} \qquad f(t) = ae^{ct} \tag{19}$$

$$\text{model 3+} \qquad f(t) = ae^{ct} + d \tag{20}$$

$$\text{model 4} \qquad f(t) = a\log(t + d_1) \tag{21}$$

$$\text{model 4+} \qquad f(t) = a\log(t + d_1) + d_2 \tag{22}$$

5.5.1 Model metrics and initial parameters

To begin with, the training of a linear model for selecting the initial parameters a_initial and b_initial for the power-law models (Eq. 15, 16, 17, 18) and their generalization is completed. To compute these initial parameters, we applied logs to both sides of the linear model, and by using the coefficients of that model, we obtained that a_initial was the exponent of the intercept ($a_{initial} = exp(coef(linear\_model)[1])$) and that b_initial was the other coefficient of the linear

model ($b_{initial} = coef(linear\_model)[2]$). This was done to accelerate the convergence of the specific models when training them.

Once these initial parameters were computed, the fitting of the non-linear models was performed. For all the power-law models (Eq. 15, 16, 17, 18), the previous a_initial and b_initial were used as the initial values for their a and b parameters. Although, for the generalization models (Eq. 16, 18), the selection of the initial values for the parameter d had to be investigated as well. In that case, after trying an iterative process for selecting d_initial, we decided to set $d\_initial = 1$. At the same time, we proceeded in different ways for making the algorithm converge by sacrificing some computational efficiency. More specifically, we fixed some of the control variables of $nls().control$ component. To be precise, $maxiter$ was increased to allow a higher number of iterations to convergence, and $minFactor$ was changed to a very small number (a positive numeric value specifying the minimum step-size factor allowed on any step in the iteration), where the increment was calculated with a Gauss-Newton algorithm and successfully halved until the residual sum of squares had been decreased or until the step-size factor had been reduced below this limit.

However, for the exponential models (Eq. 19 and 20) and for the logarithmic models (Eq. 21 and 22), the fitting of different linear models would have been useful for calculating the initial values for the parameters. In that way, for the exponential models a new linear model, that would capture the nature of the exponential model, should have provided us with initial values for the parameters a, c, and d, and following the same logic, a third linear model, following the nature of the logarithmic models, could have provided us with the respective parameters a and d. But this is considered as future work.

For the remaining models besides the null one, the AIC and the standard error of the regression estimate (s) were computed which is given by the formula:

$$s_{\text{model}} = \sqrt{\frac{RSS_{\text{model}}}{\text{residual(model)}}} \tag{23}$$

5.5.2 Selection of best model

Lastly, the respective models are trained and compared with the AIC, s, RSS formulas presented above. The results of those values are included in Tables: 9, 10, 11, 12, 13, 14, 15, and 16. The best model is defined as the one with the lowest AIC because the goal is minimizing the error between the curve and the model, and the AIC is a function of that error. To conclude, the final visualization of the best models is performed and presented in section 2, where empirical data and the curve for the best fit is plotted, that is, the fitted values for the best model compared with the initial values.

References

[1] Albert-László Barabási, Réka Albert, and Hawoong Jeong. "Mean-field theory for scale-free random networks". In: *Physica A: Statistical Mechanics and its Applications* 272.1-2 (1999), pp. 173–187.