doi:10.20312/dim.2020.01

Examples for an extended Barabási-Albert model with random initial degrees

Sándor Zsuppán
Berzsenyi Dániel Evangélikus (Líceum) Gimnázium és Kollégium
zsuppans@gmail.com

ÖSSZEFOGLALÓ. Véletlen gráfok kiterjesztett Barabási-Albert modelljének, a Cooper-Frieze modellnek három speciális esetét vizsgáljuk az egyenletes eloszlás, a Zipf eloszlás és a binomiális eloszlás alkalmazásával. A kapott elméleti összefüggéseket számítógéppel generált gráfokkal illetve a szakirodalomban fellelhető három valós hálózattal is összehasonlítjuk.

ABSTRACT. We investigate three examples for an extended Barabási-Albert model with random inicial degrees. We evaluate the general theoretical model due to Cooper-Frieze using the uniform, the Zipf and also the binomial distributions for the initial degrees of the nodes. We compare the evaluated formulae to computer-generated graphs and also to some known reference networks.

1. Introduction

The Barabási-Albert model [1] is an algorithm for generating random graphs using preferential attachment. It is an important model for producing scale-free networks, which degree distribution obeys a power-law. It serves also as a basis for many generalizations [2]. Particularly general ones of them are the Cooper-Frieze model [3] and its extension [5]. They utilize an attachement rule, which is a mixture of the preferential rule and a uniformly at random rule. They incorporate the Barabási-Albert model and many other related models as special cases.

In this short note we also investigate a special case of the Cooper-Frieze model. We describe the notation and formulate some theoretical results based on the references [3] and [5]. The main contribution of this note to the topic is the utilization of the general formulae in [3, 5] for three special cases and an illustration of the theoretical formulae with examples using NetworkX [4]. We also compare the considered cases to real networks from [6] and [7].

2. Main results

2.1. Notation and preliminaries

According to the Barabási-Albert (BA) model the undirected random graph grows by adding in each step a single node and a specified amount m of edges from this node. The terminal nodes of these edges are chosen from the set of the existent nodes according to the preferential attachement rule, i.e. with probability depending on the degree of these nodes [1, Section 5]. This process starts with a connected simple seed graph having at least m nodes

and runs until the graph achieves a prescribed amount $n(>m)$ of nodes. The resulting graph depends on the parameters n, m and on the seed. In this note we use an empty graph of m nodes and the first node connects an edge to each of them ensuring connectivity.

We investigate a variant of the BA-model: the amount of the edges from each new node to the graph is determined from the set $\{1,2,\dots,m\}$ according to a prescribed probability distribution $\boldsymbol{p}=(p_1,p_2,\dots,p_m)$. Although it is a generalization of the BA-model, it is only a very special case of the Cooper-Frieze model, see [3, Section 2.1] for its description. It is also a special case of the PARID model derived in [5] because we use a finite distribution $\boldsymbol{p}$. By [3, Theorem 1] the expected portion of nodes of degree k is well approximated by an auxiliary sequence d_k, see equations (2) and (14) in [3]. More precisely for the case considered here, the portion of nodes of degree k is a random variable δ_k the expected value of which fulfills almost surely

$$|\boldsymbol{E}(\delta_k)-\ d_k| \le C\,\frac{\log n}{\sqrt{n}} \tag{1}$$

with some positive constant C for each $k,n\ge 1$, see [3, Theorem 1]. That is, the expected portion of nodes with degree k is concentrated around the quantity d_k. Moreover, in our simplified case for $k>m$ the quantity d_k obeys a power-law with exponent 3, i.e. $d_k\sim\frac{\text{constant}}{k^3}$, see [3, Theorem 2. case (iii)]. That is, the BA-model variant considered here has the same power law exponent as the original BA-model [1,2].

The sequence d_k is defined by $d_0=0$ and

$$d_k=\tfrac{k-1}{k+2}d_{k-1}+\tfrac{2}{k+2}p_k \text{ for } k\ge 1. \tag{2}$$

Considering that $p_k=0$ for $k>m$, the equation (2) simplifies to

$$d_k=\tfrac{k-1}{k+2}d_{k-1} \text{ for } k>m, \tag{3}$$

which is equation (5.37) in [1] for the BA-model.

Remark 1. If we set the vector $\boldsymbol{p}$ for the prescribed probabilities $\boldsymbol{p}=(0,\dots,0,p_m=1)$, then the system (2) simplifies to $d_m=\frac{2}{m+2}$, which is (5.38) in [1]. Moreover, we also have $d_k=0$ for $k<m$. Hence for $\boldsymbol{p}=(0,\dots,0,p_m=1)$ we have the original BA-model.

Theorem 1. The system (2) for $1\le k\le m$ with initial condition $d_0=0$ has the unique solution

$$d_k=2\cdot\textstyle\sum_{j=1}^{k}\frac{(j+1)j}{(k+2)(k+1)k}p_j \text{ for } 1\le k\le m, \tag{4}$$

and

$$d_k=\tfrac{(m+2)(m+1)m}{(k+2)(k+1)k}\cdot d_m \text{ for } m<k. \tag{5}$$

PROOF. Substituting (4) into (2) we obtain for $1\le k\le m$ by elementary calculations that

$$\frac{k-1}{k+2}d_{k-1}+\frac{2}{k+2}p_k = 2\cdot\frac{k-1}{k+2}\sum_{j=1}^{k-1}\frac{(j+1)j}{(k+1)k(k-1)}p_j+\frac{2}{k+2}p_k$$
$$= \sum_{j=1}^{k-1}\frac{2(j+1)j}{(k+2)(k+1)k}p_j+\frac{2(k+1)k}{(k+2)(k+1)k}p_k = d_k.$$

Similarly there follows for $m < k$ that

$$\frac{k-1}{k+2}d_{k-1} = \frac{k-1}{k+2}\cdot\frac{(m+2)(m+1)m}{(k+1)k(k-1)}\cdot d_m = d_k. \blacksquare$$

Theorem 1. does not contain anything novel, we included its short proof only for the convenience of the reader. The linear recurrence (2)-(3) and its solution (4) and (5) are special cases of recurrence (1.4) and its solution (1.5) in [5], respectively. There are, however, two differences between them. Here we use a distribution vector $\boldsymbol{p}$ with only finitely many nonzero elements, and our seed graph has $m+1$ initial nodes not only two. Hence the system (2)-(3) and also its solution (4)-(5) consists of two parts: the first part (2)-(4) describes those part of the graph with nodes of degrees between 1 and m, while the second part (3)-(5) describes that with nodes of higher degree than parameter m. We investigate those two parts in the following using three selected distributions for $\boldsymbol{p}$, first a discrete uniform in Example 1. below, second in Example 2 a discrete Zipf-distribution and finally in Example 3 a binomial distribution. We compare the theoretical predictions (4)-(5) to numerically generated graphs. We also compare them to three known benchmark networks from [6] and [7].

There can be many similarities and differences between the two mentioned parts of the graph. However, here we focus mainly on one of them, which is the portion of the nodes in each of the two parts. Adding the equations (2) for $1 \le k \le m$ and using that $\boldsymbol{p}$ describes a discrete probability distribution, i.e. $\sum_{j=0}^{m} p_j = 1$, leads to

$$\sum_{k=1}^{m} d_k = 1-\frac{m}{2}d_m.$$

Setting $k = m$ in (4) and substituting it into this equation leads to

$$\sum_{k=1}^{m} d_k = 1-\sum_{j=1}^{m}\frac{(j+1)j}{(m+2)(m+1)}p_j. \tag{6}$$

By (1) there follows

$$\left|\boldsymbol{E}\left(\sum_{k=1}^{m}\delta_k\right)-\sum_{k=1}^{m}d_k\right| \le \sum_{k=1}^{m}\left|\boldsymbol{E}(\delta_k)-d_k\right| \le C\cdot\frac{m\log n}{\sqrt{n}}. \tag{7}$$

Hence, by (6) and (7) the expected portion of the nodes in the degree range $1 \le k \le m$ depends mainly on the parameter m and on the distribution $\boldsymbol{p}$. The quantity (6) can be evaluated using the generating polynomial $P(x) = \sum_{j=1}^{m} p_j x^{j-1}$ of the distribution $\boldsymbol{p}$. By $\left(x^2P(x)\right)'' = \sum_{j=1}^{m}(j+1)jp_jx^{j-1}$ equation (6) becomes

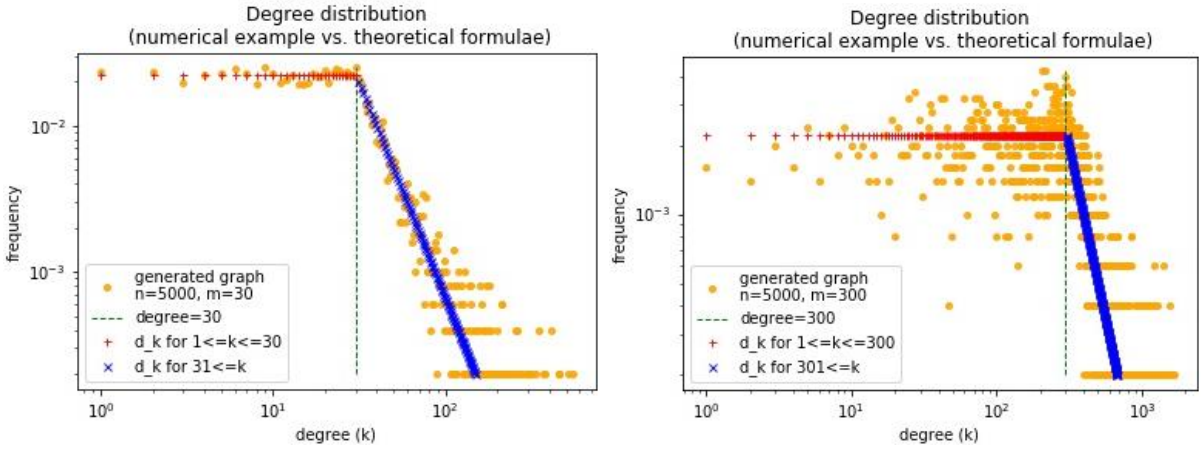
$$\sum_{k=1}^{m} d_k = 1-\frac{1}{(m+2)(m+1)}\left(x^2P(x)\right)''\Big|_{\cdot x=1}.$$

2.2. Numerical examples with generated graphs

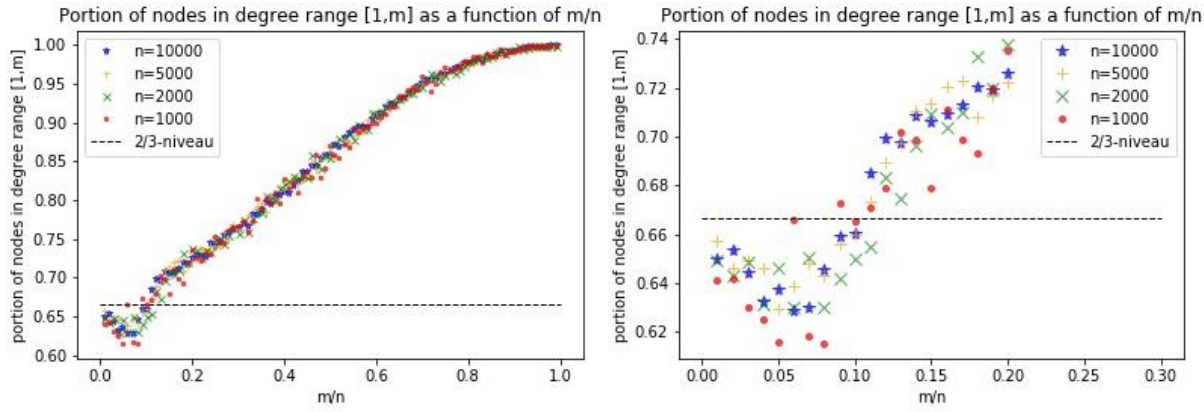
Example 1. If we set $p_j = \frac{1}{m}$ for each $j \in \{1, \dots, m\}$, i.e. the random variable describing the amount of edges each new node connects to the graph is uniformly distributed, then from (4) there follows for $1 \leq k \leq m$ that

$$d_k = 2 \cdot \sum_{j=1}^{k} \frac{(j+1)j}{(k+2)(k+1)k} \cdot \frac{1}{m} = \frac{2}{3m},$$

and for $m < k$ we have $d_k = \frac{2}{3} \cdot \frac{(m+2)(m+1)}{(k+2)(k+1)k}$. Moreover, there also follows $\sum_{k=1}^{m} d_k = \frac{2}{3}$, i.e. the portion of the nodes in degree range $\{1, \dots, m\}$ is constant. The two pictures below show a comparison of these formulae with generated graphs of $n = 5000$ nodes and parameter values of $m = 30$ and $m = 300$, respectively. For the smaller value of m we can observe a better matching of the theoretical formulae with the example graph data.



In order to understand the observable difference between the two degree distributions we plot the portion of nodes in the degree range $k \in [1, m]$ as a function of $\frac{m}{n}$ for more generated graphs.



The theoretical result, that $\sum_{k=1}^{m} d_k = \frac{2}{3}$ for uniform $\boldsymbol{p}$, holds with good approximation in case $1 \leq \frac{m}{n} \lesssim 0.1$. Above this niveau $\sum_{k=1}^{m} d_k$ grows approximately linear as a function of $\frac{m}{n}$ until $\frac{m}{n} \sim 0.7$. The slope of this linear growth was in the numerical experiments approximately 0.46

independently of n. For $0.7 \lesssim \frac{m}{n}$ the portion of the nodes saturates to 1 because the generated graph comes very near to the seed graph which for $m = n - 1$ is a star with one node of degree m and $n - 1$ nodes of degree 1.

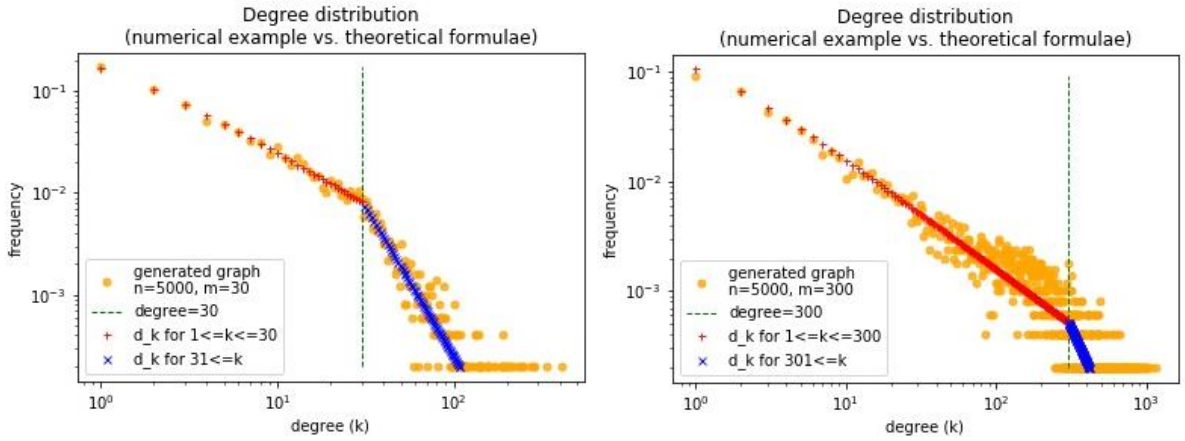
Example 2. If we set $p_j = \frac{1}{jH_m}$ for $1 \le j \le m$ in $\boldsymbol{p}$, where $H_m = \sum_{j=1}^{m} \frac{1}{j}$ denotes the m-th harmonic number, then the amount of edges each new node connects to the graph follows a discrete Zipf distribution. In this case there follows from (4) and (5) that

$$d_k = 2 \cdot \sum_{j=1}^{k} \frac{(j+1)j}{(k+2)(k+1)k} \cdot \frac{1}{jH_m} = \frac{(k+3)k}{(k+2)(k+1)} \cdot \frac{1}{H_m} \text{ for } 1 \le k \le m \text{ and}$$

$$d_k = \frac{(m+2)(m+1)m}{(k+2)(k+1)k} \cdot d_m = \frac{(m+3)m}{(k+2)(k+1)k} \cdot \frac{1}{H_m} \text{ for } m < k.$$

The expected degree distribution in the degree range $1 \le k \le m$ follows approximately a power-law with exponent 1 because $d_k \sim \frac{1}{k}$. In other words d_k itself has approximately a discrete Zipf distribution over the set $\{1, \dots, m\}$ like $\boldsymbol{p}$.

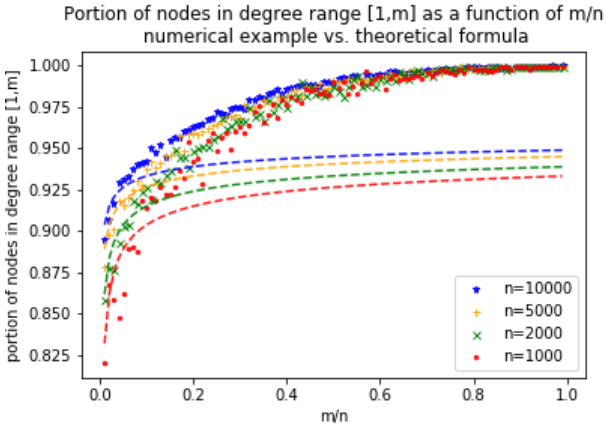
We compare these theoretical results again to two generated graphs with $n = 5000$ nodes and parameter values $m = 30$ and $m = 300$, respectively. We can observe a better matching again for the smaller parameter value as in Example 1.

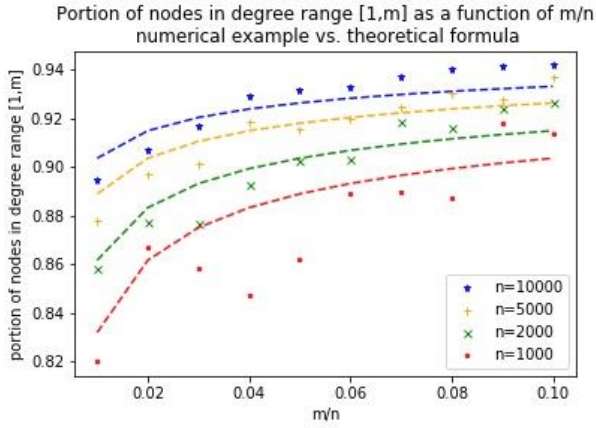


By (6) the expected portion of the nodes in the degree range $1 \le k \le m$ is

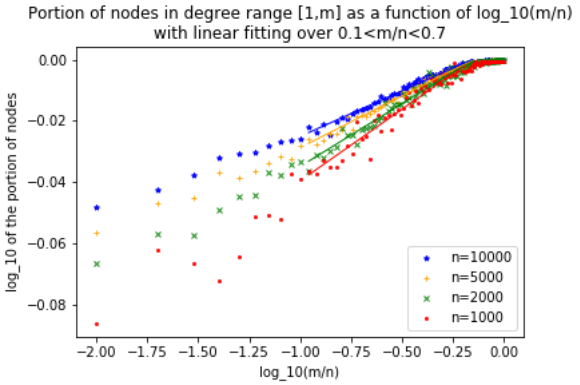
$$\sum_{k=1}^{m} d_k = 1 - \sum_{j=1}^{m} \frac{(j+1)j}{(m+2)(m+1)} \cdot \frac{1}{jH_m} = 1 - \frac{m(m+3)}{2H_m(m+1)(m+2)} \ge \frac{2}{3}, \qquad (8)$$

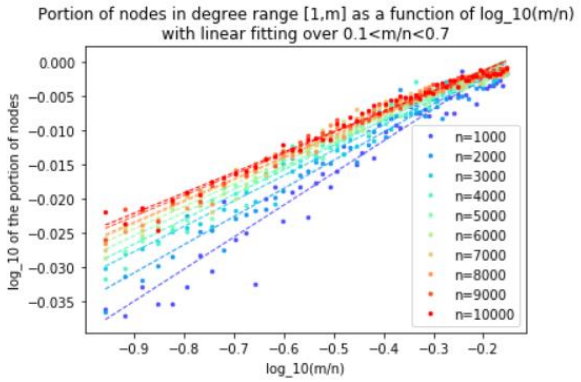
which now depends on the parameter m. We can observe this dependence on the next two diagrams, where the expected portion (8) is shown for four graph examples.



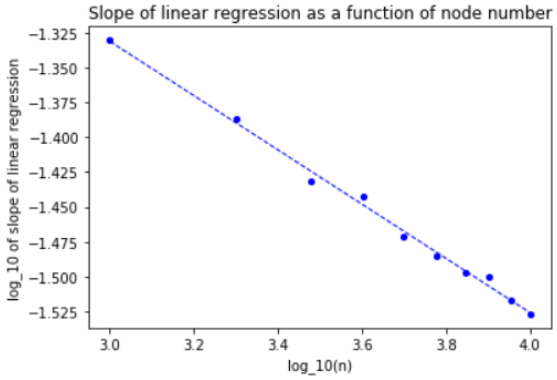


The theoretical result (8) plotted with dashed curves holds again with good approximation only in the range $1 \leq \frac{m}{n} \lesssim 0.1$. For $\frac{m}{n} \gtrsim 0.1$ the quantity $\log \sum_{k=1}^{m} d_k$ grows approximately linear as a function of $\log \frac{m}{n}$ until $\frac{m}{n} \sim 0.7$, then it saturates similar to the case in Example 1.





According to the next diagram now the slope of this growth depends on the number of nodes.



Remark 2. If the initial degree distribution $\boldsymbol{p}$ itself obeys a power-law like in Example 2., then we have to competing power-laws during the graph generating process. One of them is that discribing $\boldsymbol{p}$ and the other one induced by the preferential attachement model. According to [5, Proposition 1.3], if the power-law exponent of $\boldsymbol{p}$ is bigger than two, then that power-law with lower exponent (the more heavy tailed distribution) dominates, i.e. it will be the exponent of the power-law describing the expected degree distribution of the generated graph. Something similar happens in our Example 2. We have a Zipf distribution with exponent 1 for $\boldsymbol{p}$ and a

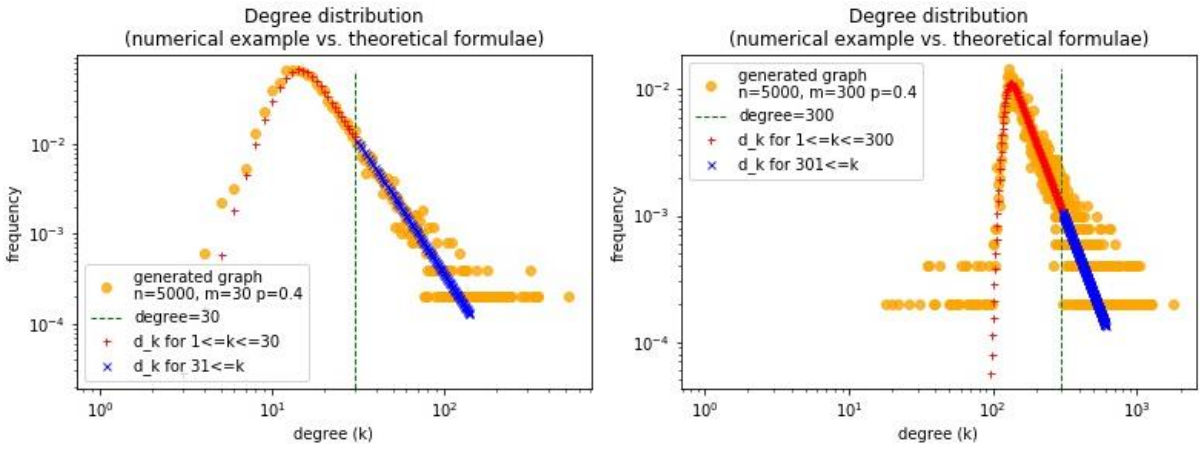
Barabási-Albert type preferential attachement model with exponent 3. However, in our example we use a finitly supported Zipf distribution and neither power-law dominates the degree distribution of the whole graph. Instead, the graph consits of two parts and in each of them one of the power-laws dominates. In the lower part, i.e. in degree range $\{1, \dots, m\}$, the exponent of the initial degree distribution dominates, while in the upper part the other one described by the BA-type preferential attachement model. This is the same case as in Example 1., where the degree distribution on degree range $\{1, \dots, m\}$ is dominated by the uniform distribution of $\boldsymbol{p}$, which itself can be interpreted as a power-law with exponent 0.

Remark 3. Although estimation (7) holds for each fixed value of m and is useful in case $n \to \infty$, it is not very informative in case when m is comparable to n. As we can observe it in both of Examples 1 and 2, there is a significant discrepancy regarding the portion of the nodes in the degree range $\{1, \dots, m\}$ between the theoretical formulae and the computationally generated graphs. However, this does not occur for relatively small parameters m. The slopes of the linear regressions above seem to depend on the number of nodes n in the experiment.

Example 3. If we set $p_j = \binom{m-1}{j-1} p^{j-1}(1-p)^{m-j}$ for $1 \leq j \leq m$ in $\boldsymbol{p}$, then the amount of initial edges from each new node follows a binomial distribution with success probability p. In this case the quantites (4) and (5) can only be evaluated numerical. However, the portion of the nodes in degree range $\{1, \dots, m\}$ can be evaluated analytically using the generating polynomial of $\boldsymbol{p}$. It becomes

$$\sum_{k=1}^{m} d_k = 1 - \frac{(m-1)(m-2)p^2+4(m-1)p+2}{(m+2)(m+1)}.$$

We compare these theoretical results again to two generated graphs with $n = 5000$ nodes and parameter values $m = 30$ and $m = 300$, respectively. Although we have now an additional parameter, namely the success probability p.



We can observe a better matching again for the smaller parameter value as in the Examples 1 and 2. However we now have an additional discrepancy between the generated and the theoretically predicted data for lower degrees.

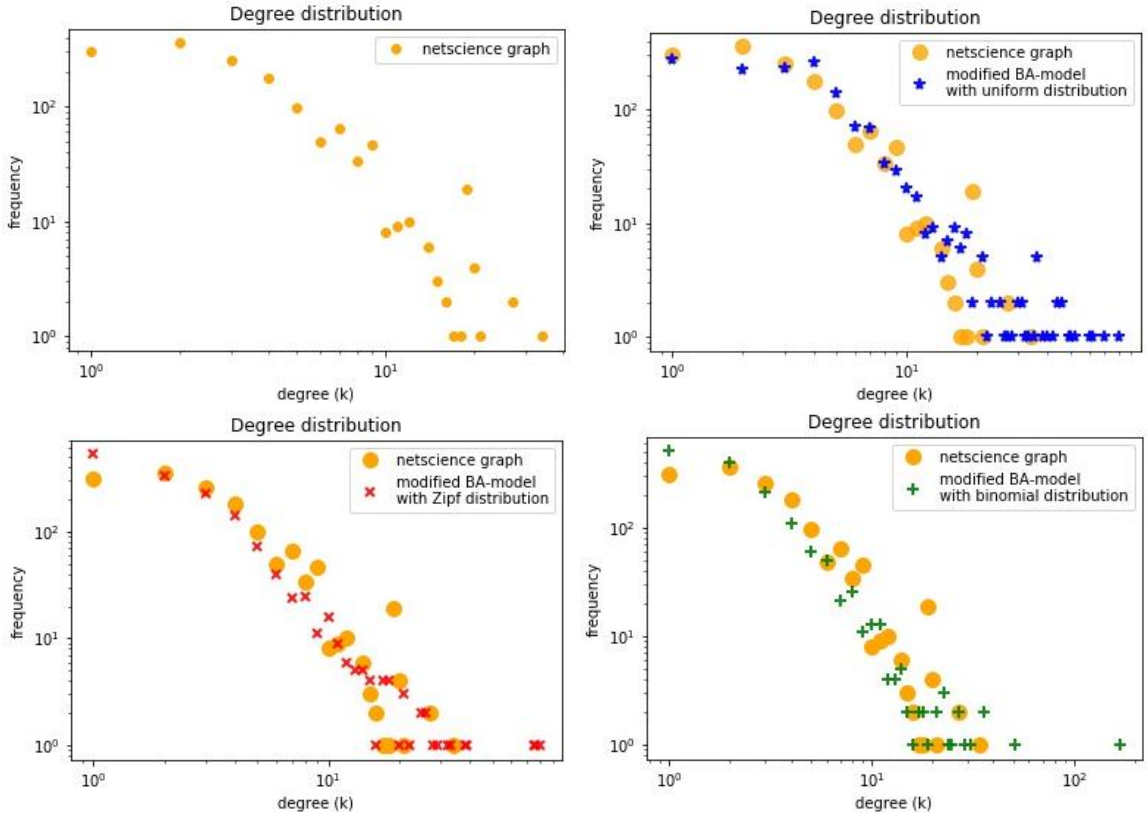
2.3. Comparison of the results with real networks

We compare the three investigated graph generating models to real networks. In the models a random graph with n nodes is generated from a star seed graph of $m+1$ nodes. According to the distribution described by $\boldsymbol{p}$ the expected number of new edges in each step is $\sum_{j=1}^{m} jp_j$. If e denotes the number of edges in the real graph, then we use the equality

$$m + (n - m - 1) \cdot \sum_{j=1}^{m} jp_j = e \tag{9}$$

for the determination of the parameter m in the models. In case of uniform distribution (9) can be solved analytically but for the Zipf and binomial distributions we obtain m resp. (m, p) numerically. For the model using binomial ditribution there is another equation needed in order to solve (9) for both parameters m and p. We have chosen the other equation so that the resulting model with binomial distributed $\boldsymbol{p}$ predicts the portion of the nodes in degree range $\{1, \dots, m\}$ with good precision.

First we use the „netscience.gml" file containing a coauthorship network of scientists working on network theory and experiment, as compiled by M. Newman in May 2006, see reference [7]. It is a graph of $n = 1589$ nodes and $e = 2742$ edges. However, among the nodes there are 128 isolated ones, therefore we take into account only $n = 1461$ nodes for comparison with the models. In view of (9) the corresponding modified BA-model with uniform $\boldsymbol{p}$ has parameter $m = 4$, the other model with Zipf distribution has $m = 3$. The parameters of the binomial model are $m = 9$ and $p = 0.11044$. For the upper tail of the degree distribution all models predict hubs with higher degrees compared to the real graph. In the middle part of the distribution all models fit fairly well.



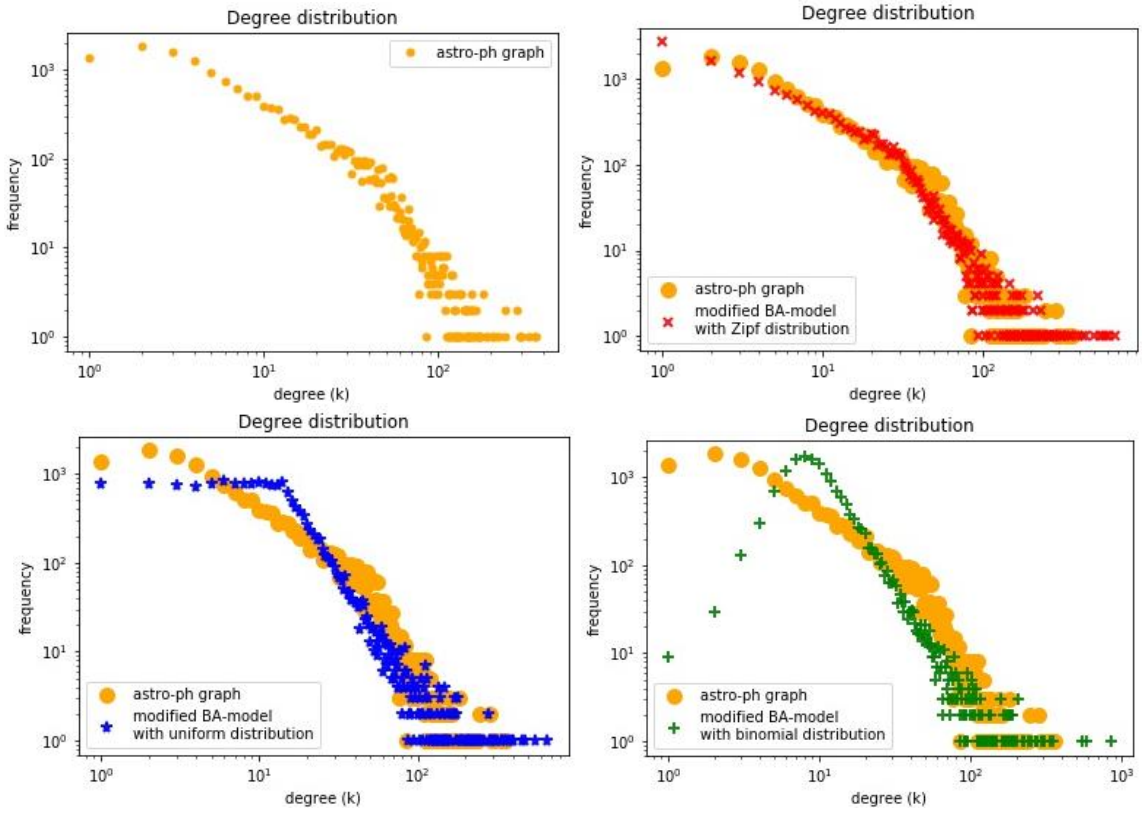
The portion of the nodes in the degree range $\{1, \dots, m\}$

- equals 0.754 in the real network for $m = 4$, which differs from the predicted portion $\frac{2}{3}$ for the uniform $\boldsymbol{p}$ with $m = 4$,
- equals 0,631 in the real network for $m = 3$, while the model with Zipf-distributed $\boldsymbol{p}$ and $m = 3$ predicts 0.755,
- equals 0,954 in the real network for $m = 9$, which correspondes good to 0,943 predicted by the binomial model. (This is of course so, because we have chosen the parameters m and p according to this.)

Hence for this network the model with binomial distributed initial degrees performs better compared to the other two models.

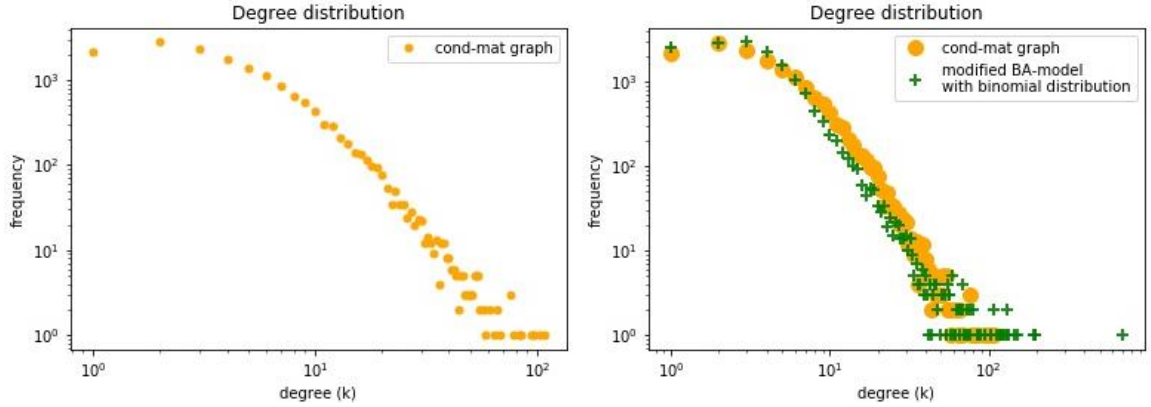
The second example is the file „astro-ph.gml" containing the collaboration network of scientists posting preprints on the astrophysics archive at www.arxiv.org, 1995-1999, as compiled by M. Newman, see reference [6]. It consists of $n = 16706$ nodes and $e = 121251$ edges. However, among the nodes there are 660 isolated ones, therefore we take into account only $n = 16046$ nodes for comparison with the models. So, the modified BA-model with uniform $\boldsymbol{p}$ has parameter $m = 14$, the model with Zipf distribution has $m = 30$ and for the model with binomial initial degree distribution we obtain $m = 13$ and $p = 0{,}5469$. The portion of the nodes in degree range $\{1, \dots, m\}$

- equals 0.689 in the real network for $m = 14$, which fits fairly good to the predicted portion $\frac{2}{3}$ for the uniform $\boldsymbol{p}$ with $m = 4$,
- equals 0,852 in the real network for $m = 30$, which is near to 0.875 predicted by the model with Zipf-distributed $\boldsymbol{p}$ and $m = 30$,
- equals 0,670 in the real network for $m = 13$, which corresponds good to 0,678 predicted by the binomial model. (But this is of course so, because we have chosen the parameters m and p according to this.)



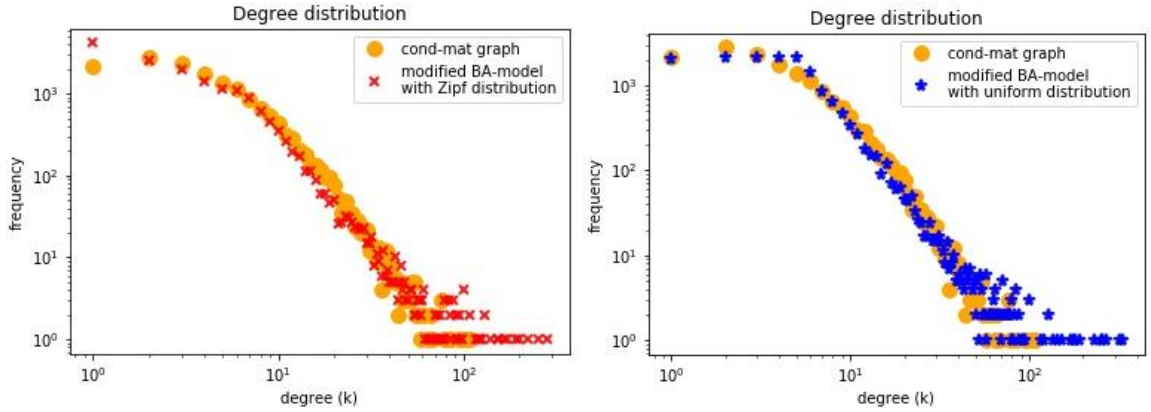
Altough the model using binomial distribution was chosen so that its expected number of nodes and its expected portion of nodes in degree $\{1, \dots, m\}$ range fit to the corresponding characteristics of the real network, its degree distribution differs considerably from that of the real network. In this example the Zipf-based model approximates the real network best.

The third example we investigate here is „cond-mat.gml" containing the collaboration network of scientists posting preprints on the condensed matter archive at www.arxiv.org, 1995-1999, as compiled by M. Newman [7]. It consists of $n = 16726$ nodes and $e = 47594$ edges. However, there are 462 isolated nodes, therefore we take into account only $n = 16264$ nodes for comparison with the models. So, the modified BA-model with uniform $\boldsymbol{p}$ has parameter $m = 5$, the model with Zipf distribution has $m = 7$ and the model with binomial distribution $m = 28$ and $p = 0{,}07148$.



Concerning the portion of nodes in degree range $\{1, \dots, m\}$:

- with $m = 5$ this portion of nodes in the real network is 0,657 which comes fairly close to the predicted $\frac{2}{3}$ by the model with uniform initial degree distribution,
- with $m = 7$ this portion of nodes in the real network is 0,769, which matches the predicted 0,813 by the model with Zipf initial degree distribution rather good,



- with $m = 28$ this portion of nodes in the real graph is 0,986 which comes close to the predicted 0,985 by the model with binomial initial degree distribution just because its parameters were constructed so.

In this example perform all three models again fairly good. However, considering the upper tail of the degree distributions, the binomial-based model performs better than the other two, because it predicts not so many hubs then the other two models.

3. Conclusion

We have investigated three examples for a Barabási-Albert type model with random initial degrees. In all examples happend the degree distribution of the resulting graph to consist of two parts: the lower tail of it was biased by the given initial degree distribution while the upper tail by the used preferential attachement model. This partition was also observable in the investigated reference networks. None of the considered models performed equally good for all real examples. The predictions of the theoretical formulae were more accurate in case the maximum initial degree is considerably less than the number of all nodes. The more the maximum initial degree increases, the more is the resulting graph influenced by the seed graph in our models.

Reference

[1] **Barabási A-L.**, Network Science, Libri Könyvkiadó, Budapest 2019. http://networksciencebook.com/
[2] **Bollobás B., Riordan O.M.**, Mathematical results on scale-free random graphs, in: Handbook of Graphs and Networks: From the Genome to the Internet, Wiley-VCH Verlag GmbH & Co. KgaA (2002), doi.org/10.1002/3527602755.ch1
[3] **Cooper C., Frieze A.**, A general model of web graphs, Random Sctructres and Algorithms Vol. 22 Issue 3 (2003), 311-335. doi.org/10.1002/rsa.10084
[4] **Hagberg A., Schult D., Swart P.,** Exploring Network Structure, Dynamics, and Function using NetworkX in Proceedings of the 7th Python in Science conference (SciPy 2008), G Varoquaux, T Vaught, J Millman (Eds.), pp. 11-15. https://networkx.github.io/documentation/stable/
[5] **Deijfen M., van den Esker H., van der Hofstad R., Hooghiemstra R.**, A preferential attachment modelwith random initial degrees, Ark. Mat.,47(2009), 41-72. doi.org/10.1007/s11512-007-0067-4
[6] **Newman M. E. J.**, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA 98, 404-409 (2001) http://www-personal.umich.edu/~mejn/netdata/
[7] **Newman M. E. J.**, Finding community structure in networks using the eigenvectors of matrices, Preprint physics/0605087 (2006) http://networkdata.ics.uci.edu/data/netscience/