Simple rules govern finite-size effects in scale-free networks

S. Cuenda¹,² and J. A. Crespo¹

¹ Departamento de Economía Cuantitativa, Facultad de CC. Económicas y Empresariales, Universidad Autónoma de Madrid - C. Fco. Tomás y Valiente 5, E-28049 Madrid, Spain, EU
² Grupo Interdisciplinar de Sistemas Complejos (GISC)

received 15 March 2011; accepted in final form 14 June 2011
published online 11 July 2011

PACS 89.75.Da – Systems obeying scaling laws
PACS 87.23.Ge – Dynamics of social systems
PACS 02.50.Cw – Probability theory

Abstract – We give an intuitive though general explanation of the finite-size effect in scale-free networks in terms of the degree distribution of the starting network. This result clarifies the relevance of the starting network in the final degree distribution. We use two different approaches: the deterministic mean-field approximation used by Barabási and Albert (but taking into account the nodes of the starting network), and the probability distribution of the degree of each node, which considers the stochastic process. Numerical simulations show that the accuracy of the predictions of the mean-field approximation depend on the contribution of the dispersion in the final distribution. The results in terms of the probability distribution of the degree of each node are very accurate when compared to numerical simulations. The analysis of the standard deviation of the degree distribution allows us to assess the influence of the starting core when fitting the model to real data.

Introduction. – Power laws are not a new issue in scientific literature. The emergence of the scale-free behavior in the degree distribution of the sizes of biological genera, incomes, words in a text, scientific citation, etc., has been widely studied (and several times re-invented) in the past century (see [1] for an interesting review of this re-inventions). One of the most famous works in this subject nowadays is the one by Barabási, Albert and Jeong (BA) [2], in which they introduce the “preferential attachment” model in social networks such as the world wide web [3] and the network of movie actors [4]. In this model, nodes would get new links from new nodes in the network with probability proportional to their degree. In this way, nodes with high degree would be more likely to receive new links, a rich gets richer mechanism that would render a power-law distribution of the network degree. This model has been widely used by other researchers (see, for example, [5] for an exhaustive review of other models constructed thereafter, and references therein).

However, some investigations showed that the model of preferential attachment of BA would depart from the predicted power-law behavior in small networks [6–9]. An influence of the initial nodes from which the network starts growing was acknowledged, but no general prediction of the effect of these nodes on the degree distribution was made, except for particular cases [6–8]. Although the attempts of studying this effect have been both numerical and theoretical, to our knowledge there are neither intuitive, general explanations of the process, nor predictions of the final degree distribution in these networks in the scientific literature yet.

In this letter we find a general, theoretical prediction of the final degree distribution of finite networks growing with preferential attachment in terms of the degree distribution of the starting network. We obtain an expression of the final distribution using two different approaches: the well-known, deterministic mean-field approximation (with the contribution of the nodes of the starting network), and the expected probability distribution of the degree of each node, which considers the stochastic process. The methods used are very simple and intuitive, and the numerical simulations support very well the theoretical results. One of our main findings is the relevance of the starting nodes of the network in the final degree distribution, which must be considered when fitting the model to real data.

Model definition. – The model on which we are going to focus is the original one introduced by BA [2]. In this...
model, at every time step, a new node arrives to the network and attaches to other nodes by \( m \) undirected new links, the probability of any node in the network of gaining one of these new links being proportional to its degree. Notice that, in order for the process to be well defined, a starting network (or core) to which the first new node may link is needed. We will not allow multiple linking between two nodes, thus the size of the starting core must be of \( m \) nodes, at least. We do not consider nodes with null degree since, in this model, these nodes would never get any links.

We define \( t \) as the number of nodes at each time step and \( t_0 \) as the number of nodes in the starting core, with \( t_0 \geq m \). If a node has degree \( k \) at time \( t \), then the probability of this node gaining a new link when node \( t + 1 \) arrives is, according to the model of BA [2], \( \pi_{k,t} = Ak \), where \( A \) is a normalizing constant that must satisfy the expected number of new links in the network to be \( m \) with \( \pi \).

\[
\pi_{k,t} = \frac{mk}{\sum_{i=1}^{t} k_i}.
\]

If the mean number of links in the starting core is \( m_0 \), then the total degree of the network at time \( t \) is \( \sum_{i=1}^{t} k_i = 2[(m_0 - m)t_0 + mt] = 2m(\mu_0 + t) \), with \( \mu = (m_0/m - 1) \). Notice that \( \pi_{k,t} = k/(2(\mu_0 + t)) \) is such that the dynamics of every node is independent from the rest of nodes in the network. Therefore, whatever the approach to simulate the dynamics of each node, we can use this result to calculate the degree distribution of the network for each \( t \geq t_0 \), knowing the initial distribution at \( t_0 \).

**Mean-field approximation.** – We start with the mean-field approach followed by BA [2], which consists in a continuum approximation in both degree and time in such a way that the rate of change of the degree of any node, \( k \), equals its expected value, \( \pi_{k,t} \):

\[
\frac{dk}{dt} = \frac{k}{2(\mu_0 + t)}.
\]

Integration of \((2)\) renders the deterministic degree at time \( t \) of a node that has degree \( \kappa \) at time \( \tau \), \( k(t) = h(t; \kappa, \tau) \), with

\[
h(t; \kappa, \tau) = \kappa \sqrt{\frac{\mu_0 + t}{\mu_0 + \tau}}.
\]

and \( t, \tau \geq t_0 \). From expression \((3)\) it follows that, for fixed \( t \) and \( \tau \), \( k(t) \) strictly increases with \( \kappa \), and for fixed \( t \) and \( \kappa \), \( k(t) \) strictly decreases with \( \tau \). Usually, \( \kappa = m \) is taken for all nodes, and the usual asymptotic, power-law behavior is obtained [2]. Notice that, however, this initial condition is only valid for the nodes added to the network. When we consider the case of the added nodes (\( \tau > t_0 \)) separately from the case of the nodes of the starting core (\( \tau = t_0 \)), the finite-size effect emerges.

Let \( F_m(k, t) \) be the complementary, cumulative distribution of the degree of the network at time \( t \) under the mean-field approximation. Thus, \( F_m(k, t) \) gives the portion of nodes at time \( t \) with degree greater than or equal to \( k \) for \( t \geq t_0 \). Let \( F_0(k) \) be the corresponding degree distribution of the starting core at time \( t_0 \), thus \( F_0_k(t) = F_0(k) \).

We define \( k_m(t) \) as the degree at time \( t \) of the nodes that had degree \( m \) at time \( t_0 \), i.e., \( k_m(t) = h(t; m, t_0) \). Therefore, all the added nodes of the network must have degree smaller than \( k_m(t) \) at time \( t \). Similarly, we define \( k_0(k, t) \) as the degree that should have a node at time \( t_0 \) in order to have degree \( k \) at time \( t \); from eq. \((3)\) it follows that \( k_0(k, t) = h(t; 0, k, t) \).

At time \( t \), nodes with degree \( k > k_m(t) \) cannot come from the added nodes, but from the nodes that in the starting core had degree \( k_0(k, t) \), instead. Thus, the portion of nodes with degree greater than or equal to \( k \), with \( k > k_m(t) \), is \( (t_0/t)F_0(k_0(k, t)) \). For \( m < k \leq k_m(t) \), the portion of nodes with degree greater than or equal to \( k \) coming from the starting core are \( (t_0/t)F_0(k_0(k, t)) \), but also added nodes until time \( \tau^* \), with \( \tau^* \) such that \( h(t; m, \tau^*) = k \), must be considered, rendering \((\tau^* - t_0)/t \). Finally, for \( k \leq m \) all added nodes have degree greater than or equal to \( k \), and the contribution from the starting core is similar to the other ranges. Therefore, the final degree distribution at time \( t \) is

\[
F_m(k, t) = \begin{cases}
1 - \frac{t}{t_0} \left[ 1 - F_0(k_0(k, t)) \right], & k \leq m, \\
\left( \frac{\mu + 1}{\mu} \right) ^2 - \mu \frac{t_0}{t} - \frac{1}{\mu} \left[ 1 - F_0(k_0(k, t)) \right], & m < k \leq k_m(t), \\
\frac{t_0}{t} F_0(k_0(k, t)), & k > k_m(t).
\end{cases}
\]

The finite-size effect that makes expression \((4)\) depart from the classical power law comes from the starting core in terms of \((t_0/t)F_0(k_0(k, t))\) and from the finite number of added nodes, which yields the emergence of \( k_m(t) \). Notice that these contributions vanish in the limit \( t \to \infty \), where the usual asymptotic result is recovered.

Previous works of the finite-size effect [6–8] studied the ratio between the actual final degree density distribution of the finite network and the asymptotic power law of BA and found that, for networks growing from the same starting core, this ratio is

\[
\frac{f(k, t)}{f(k, t \to \infty)} = w(k/\sqrt{t}),
\]

where \( w(x) \) is the cut-off function, and depends on the starting core used. From expression \((4)\) we can calculate such a ratio within the mean-field approximation, rendering

\[
\frac{f_m(k, t)}{f_m(k, t \to \infty)} = \frac{\mu_0 + t}{t} u(k, t)
\]

where \( f_m(k, t) = -\partial F_m(k, t)/\partial k \) is the density distribution in the mean-field approximation, \( f_m(k, t \to \infty) = 2m^2k^{-3} \) is the asymptotic density obtained with the BA methodology,

\[
u(k, t) = \begin{cases}
1 + g(k_0(k, t)), & m < k < k_m(t), \\
g(k_0(k, t)), & k > k_m(t),
\end{cases}
\]

38002-p2
\( g(k) = \frac{1}{2m(t+1)} k^3 f_0(k) \) and \( f_0(k) = -dF_0(k)/dk \) is the degree distribution of the starting core. Noticing that \( k_0(t, k) = k \sqrt{\mu_0 t + t_0}/(\mu_0 t + t) \), expression (6) resembles expression (5) when \( \mu_0 t/2 \leq 0 \).

Expressions (4) and (6) are obtained within the well-known mean-field approximation as a consequence of considering the starting core degree distribution in the calculation of \( F_m(k, t) \). However, it is strictly deterministic: the degree distribution of the starting \( t_0 \) nodes at time \( t \) is the same initial distribution at time \( t_0 \), stretched by a factor \( \sqrt{\mu_0 t + t_0}/(\mu_0 t + t) \); no effect of the dispersion of the degree of nodes as a consequence of stochastic expression (5) when \( t \). The cumulative distribution \( F_e(k, t) \) can be calculated from (9) using the expression

\[
F_e(k, t) = \sum_{j>k} f_e(j, t).
\]

Numerical results. – In order to check the results of expressions (4), (9) and (10) we need to simulate the stochastic process defined by expression (1). However, for \( m > 1 \), this process does not cover the space of probabilities of the whole system, i.e., the sum of the probabilities for all possible choices of the nodes that get a new link is not normalized to unity, and thus the condition

\[
\sum_{i_1=1}^t \sum_{i_2<i_1}^t \cdots \sum_{i_m<i_{m-1}}^t \pi_{k_{i_1}, t} \pi_{k_{i_2}, t} \cdots \pi_{k_{i_m}, t} = 1
\]

is not fulfilled for \( m > 1 \) (though it is for \( m = 1 \)). Therefore, for \( m > 1 \), there is no such stochastic process.

Nonetheless, we can consider the following stochastic linking process when a new node arrives: the \( m \) nodes are chosen sequentially, each one with probability \( k/\sum_i k_i \), where the sum does not contain previously chosen nodes, and avoiding repetition. Expression (1) describes exactly this stochastic process for \( m = 1 \), and it is a good approximation for \( t \gg m \). For \( t \gg m \), there is an exclusion effect that makes the probability highly non-linear with respect to the degree (the case \( t = m \), where all nodes should get a new link with probability equal to 1, shows the inaccuracy), and the model does not describe well the growing process in this regime. However, as the network grows, the model will eventually capture the stochastic dynamics of the nodes. As a result, for \( t_0 = m > 1 \), the dynamics of young nodes are well approximated by expression (1), but the dynamics of old nodes may depart from that. These effects render a slight error in the prediction of the final distribution \( F(k, t) \) for large values of \( k \).

Figures 1 and 2 show the agreement for different starting cores between the mean-field approximation \( F_m(k, t) \), the expected \( F_e(k, t) \) and synthetically generated complementary cumulative distributions, \( F_{\text{syn}}(k, t) \). The growing process of fig. 1(a) has \( m = 1 \), and therefore the result given by \( F_e(k, t) \) is exact in this case. However, it is also the worst example for the mean-field approximation, \( F_m(k, t) \), since the tail of the distribution comes from the dispersion of the degree of the starting nodes in the network; this is clear when comparing the tail of the distribution of the averaged simulations, \( \langle F_{\text{syn}}(k, t) \rangle \), with one of them, \( F_{\text{syn}}(k, t) \). In fig. 1(b) there is less dispersion, and the mean-field approximation works better. The expected distribution \( F_e(k, t) \) departs slightly in the tail from \( \langle F_{\text{syn}}(k, t) \rangle \) since the starting number of nodes is \( t_0 = 10 \) and the number of links per new node is \( m = 10 \),
S. Cuenda and J. A. Crespo

but the result is not so bad since, as we add nodes to the network, the model improves its accuracy.

In fig. 2 we see the effect in the final distribution of a poorly connected (fig. 2(a)) and a highly connected (fig. 2(b)) starting core compared to the number of links per new node, $m$. The heap that emerges in the tail of the latter comes from the degree of the starting nodes, higher than the initial degree of the added nodes. In
The methodology followed in this letter can also be applied to other growing network models where the dynamics of the degree of a single node can be well approximated by its own state, regardless of the state of other nodes or the degree distribution of the network. These models can lead in the asymptotic limit to power-law–tailed distributions (see [5]) that are fitted against real data which is supposed to be modelled by this kind of growing mechanisms. However, the shape of the final distributions of these models may depend strongly on the initial configuration used, even for large networks, as shown in fig. 3, where the final network is.
10^2 and 10^3 times larger than the starting core. Clearly, the fitting results may differ significantly depending on the size and distributions of the initial cores of the models, and therefore the results presented in this letter should be considered.

Nowadays, dynamic networks are the hot topic in network investigation. Reaction kinetics on metabolic networks, spread of information or viruses in social networks... are examples of dominant issues in the latest scientific literature. But, quoting Barabási, “to make progress in this direction, we need to tackle the next frontier, which is to understand the dynamics of the processes that take place on networks” [10]. We hope that this work may help in that understanding.

***

SC is partially supported by MEC grant FIS2006-01485. JAC is partially supported by MEC grant ECO2010-19596.

REFERENCES

[1] Simkin M. V. and Roychowdhury V. P., *Phys. Rep.*, 502 (2011) 1.
[2] Barabási A.-L., Albert R. and Jeong H., *Physica A*, 272 (1999) 173.
[3] Barabási A.-L., Albert R., Jeong H. and Bianconi G., *Science*, 287 (2000) 2115.
[4] Albert R. and Barabási A.-L., *Phys. Rev. Lett.*, 84 (2000) 5660.
[5] Newman M. E. J., *SIAM Rev.*, 45 (2003) 167.
[6] Dorogovtsev S. N. and Mendes J. F. F., *Phys. Rev. E*, 63 (2001) 62101.
[7] Krapivsky P. L. and Redner S., *J. Phys. A: Math. Gen.*, 35 (2002) 9517.
[8] Waclaw B. and Sokolov I. M., *Phys. Rev. E*, 75 (2007) 056114.
[9] Guimarães P. R., de Aguiar M. A. M., Bascompte J., Jordano P. and dos Reis S. F., *Phys. Rev. E*, 71 (2005) 37101.
[10] Barabási A.-L., *Science*, 325 (2009) 412.