Eur. Phys. J. B **61**, 377–388 (2008)
DOI: 10.1140/epjb/e2008-00075-3

THE EUROPEAN
PHYSICAL JOURNAL B

Growing directed networks: stationary in-degree probability for arbitrary out-degree one

D. Fraiman[a]

Departamento de Matemática y Ciencias, Universidad de San Andrés, Buenos Aires, Argentina

Received 29 July 2007 / Received in final form 26 December 2007
Published online 22 February 2008 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2008

Abstract. We compute the stationary in-degree probability, $P(k_{in})$, for a growing network model with directed edges and arbitrary out-degree probability. In particular, under preferential linking, we find that if the nodes have a light tail (finite variance) out-degree distribution, then the corresponding in-degree one behaves as k_{in}^{-3}. Moreover, for an out-degree distribution with a scale invariant tail, $P(k_{out}) \sim k_{out}^{-\alpha}$, the corresponding in-degree distribution has exactly the same asymptotic behavior only if $2 < \alpha < 3$ (infinite variance). Similar results are obtained when attractiveness is included. We also present some results on descriptive statistics measures such as the correlation between the number of in-going links, K_{in}, and outgoing links, K_{out}, and the conditional expectation of K_{in} given K_{out}, and we calculate these measures for the WWW network. Finally, we present an application to the scientific publications network. The results presented here can explain the tail behavior of in/out-degree distribution observed in many real networks.

PACS. 05.65.+b Self-organized systems – 89.75.Kd Patterns – 87.23.Ge Dynamics of social systems – 02.50.Cw Probability theory

1 Introduction

Barabási and Albert [1] discovered that several networks in nature have a strange topological characteristic: they have a scale-free [2–4] degree distribution, $P(k) \sim k^{-\alpha}$, where the degree of a vertex is defined as it total number of connections. Nowadays, this empirical behavior is confirmed in a great number of completely different empirical networks, from biological networks to e-mail networks, including scientific publication networks. Focusing on undirected networks, in [1] the authors also proposed a model (B-A model) for explaining this behavior. The model can be formulated as follows: 1) start with a network with N nodes, connected by j edges in an arbitrary way; and 2) at each time step, a new node with m edges appears, each edge connecting to the existing nodes according to some probability law, π. The probability that a new edge attaches to a node with degree k, π^k, was defined [1] as proportional to the degree of the node. In particular, they showed that with this attachment law,

$$\pi^k = \frac{kN^k}{\sum\limits_{j\in\mathbb{N}} jN^j}, \qquad (1)$$

where N^k is the number of nodes with degree k, the stationary degree distribution has a power law tail, $P(k) \sim k^{-3}$. In [5] they computed the stationary degree probability (not only the tail behavior) for the B-A model, but for a generalization of the preferential linking attachment law. They introduced a new parameter, the attractiveness, A (in their case $A \geq -m$), and defined the attachment law as:

$$\pi^k = \frac{(k+A)N^k}{\sum\limits_{j\in\mathbb{N}} (j+A)N^j}. \qquad (2)$$

In this case, they found that $P(k) \sim k^{-(2+A/m)}$, being more flexible for comparing to empirical networks. Typically, degree distribution of real networks satisfy, $P(k) \sim k^{-\alpha}$ with $2 \leq \alpha \leq 3$. But the B-A model, no matter which is the attachment law, has a major drawback, the number (m) of edges that arise from new nodes is a fixed number. In almost all real networks, the new nodes do not have the same number of edges. On the other hand, the number of edges of a random selected new node (from a real network) is a random variable. So, in order to be more realistic, we will study the B-A model in the case where the new nodes appear with a random number of edges, but in the more general context of directed growing networks. In this context, new questions arise from empirical networks.

[a] e-mail: dfraiman@udesa.edu.ar

Directed networks are characterized by the fact that the edges are directed (arrows), each node has edges that point at it, and others that born in it. The in-degree of a node is defined as the number of incoming edges, the out-degree as the number of its outgoing edges, and the degree is the sum of the two previous ones. The most studied directed growing networks have been the WWW network [7,8,13], and the scientific publications network [6]. In the first one, each node represents a web page and the hyper-links (references to other web pages) represents the directed edges or links. In the second one, each paper is a node, and its references the directed links. In this last case, the in-degree distribution represents the distribution of citations for a random selected paper, and the out-degree distribution represents the number of references of a random selected paper. Interestingly, real directed growing networks follow in general one of two possible behaviors. In the first case they have an out-degree exponential distribution, $P(k_{out}) \sim a^{k_{out}}$ $(0 < a < 1)$, or an out-degree distribution taking finitely many values, associated with an in-degree distribution with a power law tail $P(k_{in}) \sim k_{in}^{-\alpha}$ where typically $\alpha \approx 3$. In the second case the out-degree distribution satisfies $P(k_{out}) \sim k_{out}^{-\beta}$, and is associated with $P(k_{in}) \sim k_{in}^{-\alpha}$ with $\alpha \approx \beta$. Examples, such as biological, WWW, or communication networks, can be found in [2–4,9].

In this paper, we address the question of why the empirical growing directed networks show this strange general behavior for the tail of the in/out degree distributions. We study a particular growing network model (a generalization of the B-A model to be precise), obtaining the stationary joint in-out degree distribution, $P(k_{in}, k_{out})$, and some of its derivatives, such as the marginal distribution, $P(k_{in})$, the covariance, and the conditional expectation of the number of in-links given the number of out-links. In particular, studying in detail $P(k_{in})$, we prove (for the model presented here) that it is expected to observe the in/out tail behavior reported for real networks [2–4]. Finally we present an application to the most “pure” (extremely few double arrows) growing directed network: the scientific publication network. In this application, we show the relevance of having an expression for the limit in-degree distribution ($P(k_{in})$) for an arbitrary out-degree one ($P(k_{out})$).

2 Growing directed network model

Before describing the model, it is important to remark that real directed growing networks have in general a considerable asymmetry between the in-links and out-links of a node. For example, nobody will care much about how many references (out-links) an own paper has, but people are interested in the number of cites (in-links) that their own paper has. That is why we are going to treat the out-links from a new node and the in-links in a completely different way. In particular, a node can receive (with positive probability), a connection from a new node at any moment, but typically a node can not change who their pointers (the set of nodes it is pointing to) are. This is very clear in the scientific publications network. In this network the in-degree distribution has been extensively study [6,8], whereas the out-degree distribution has been poorly reported [10,12]. Nevertheless, in the case of the WWW network, the outgoing links (hyper-links) can change at any moment and new hyper-links can be aggregated or old hyper-links can be redirected. In [7,8] they proposed some models for describing this network taking into account the characteristics mentioned above. However these models do not consider that the new nodes have a particular out-degree distribution, i.e. the models are constructed under the hypothesis that new nodes have a fixed number of out-links. The major problem of both models is that the nodes (web-pages) do not have a controlled number of out-links, they can have a huge number of them which does not seem realistic. Our strategy for modeling these networks is completely different to the ones proposed in [7,8]. For us, most of the variability in the number of out-links is explained when the node appears, defined as “intrinsic” variability, and not as a product of updating nodes. We think that in many real networks the updating of nodes can give a small correction compared with the “intrinsic” variability. This assumption is at the core of our model. In a real network the “intrinsic” variability is given by different reasons that are hard to know (why does a randomly selected scientific paper has a number of references with some particular distribution?), but typically the problem of trying to understand this variability is not a major question.

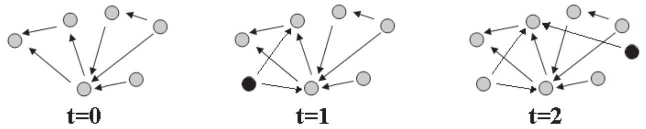

Fig. 1. Scheme of the growing network model. In each temporal step a new node (shown in black) with K_{out} out-links appears; these links point towards existing nodes. K_{out} is not a fixed number, on the contrary it is a random variable. The degree vector at time 0, and 1 is: $\boldsymbol{N}_0 = (1,4,0,0,1,0,0,0,...,0,...)$, $\boldsymbol{N}_1 = (1,4,1,0,0,1,0,0,...,0,...)$.

We define $\boldsymbol{N}_n = (N_n^1, N_n^2, ..., N_n^k, ...)$ as the degree vector at time n, where N_n^j is the number of nodes with degree j at time n, and $\boldsymbol{N}_{in,n} = (N_{in,n}^1, N_{in,n}^2, ..., N_{in,n}^k, ...)$ as the in-degree vector, where $N_{in,n}^j$ is the number of nodes with in-degree equal j at time n. Now, we describe the growing network model: 1) initially the network consists of M nodes connected in a given arbitrary way; 2) at each time step, say time step $n+1$, a node with K_{out} outgoing-edges appears, where K_{out} is a random variable ($\sum_{k_{out}\in\mathbb{N}} P(K_{out} = k_{out}) = 1$); and 3) each new directed edge points out to an existing node with some probability law π_{n+1} (uniform, preferential linking, etc.). Figure 1 illustrates a scheme of the model. If π_{n+1} is an arbitrary function that depends on $\boldsymbol{N}_n$, and/or $\boldsymbol{N}_{in,n}$ ($\boldsymbol{N}_{out,n}$), then the growing network model, described above is a Markov chain taking values in $\mathbb{N}_o^{\mathbb{N}}$ ($\mathbb{N}_o = \mathbb{N} \cup \{0\}$) or

$\mathbb{N}_o \times \mathbb{N}_o^{\mathbb{N}^2}$ with transition probabilities given by π_{n+1}. In this work (under the Markovian hypothesis), we show an easy way to compute stationary (in/out) degree probabilities for arbitrary π_{n+1}. An important part of this article is devoted to the study of the model under the law:

$$\pi_{n+1}^k = \frac{(A+k)N_n^k}{\sum_{j\in\mathbb{N}}(A+j)N_n^j}, \tag{3}$$

and in Section 2.4 we show some results under different π's. The law of equation (3) corresponds to preferential linking on degree with attractiveness. This probability is well defined for values of A greater or equal to $-B$, where

$$B := \min_j\{j : P(K_{out} = j) > 0\}. \tag{4}$$

For this attachment law, the model is in fact an extension of the Albert-Barabási model, although in this case K_{out} is a random variable with an arbitrary distribution, $P(K_{out} = k_{out})$ with $k_{out} \in \mathbb{N}$, and the edges are directed. The limit (stationary) in-degree distribution and the limit degree distribution have not been reported, even for simple cases as K_{out} taking values 1 and 2, with probabilities p_1 and $1 - p_1$ respectively. Moreover, even in the undirected case, it is not known if in general the limit degree distribution $(P(k))$ satisfies a superposition principle (linear combination).

2.1 Stationary probabilities

The number of out-links does not depend on time (see Sect. 2.5 for additional details), therefore, the limit out-degree distribution satisfies $P(k_{out}) \equiv P(K_{out} = k_{out})$. Note that the out-degree distribution is defined a priori (in accordance with the specific network), imposing in this way the asymmetry mentioned before between the in and out links of a node. We are interested in obtaining the limit degree distribution, $P(k)$, and the limit in-degree one, $P(k_{in})$. In order to compute this last probability function, we first compute the stationary joint degree and out-degree distribution, $\mathcal{P}(k, k_{out}) := P(K = k, K_{out} = k_{out})$, where K is the degree $(K = K_{in} + K_{out})$ of a random selected node. If the network is distributed according to the stationary probability, then the probability that a randomly chosen node has k_{out} out-links and k total links, $\boldsymbol{K} = (K, K_{out}) = (k, k_{out})$, is given by:

$$\mathcal{P}(k, k_{out}) = P(\boldsymbol{K} = (k, k_{out})) = \lim_{n\to\infty} \frac{\mathcal{N}_n^{k,k_{out}}}{\sum_{h,i\in\mathbb{N}} \mathcal{N}_n^{h,i}}$$

where $\mathcal{N}_n^{h,i}$ is the number of nodes with h total links from which i are out-links at time n. The last equality holds by the Law of Large Numbers for Markov chains. Clearly, the joint in-out degree can be computed from this last one, $P(k_{in}, k_{out}) = \mathcal{P}(k_{in} + k_{out}, k_{out})$, and also the in-degree and degree probability taking marginal distributions.

$\mathcal{N}_{n+1}^{j,k}$ depends on: 1) $\mathcal{N}_n^{j,k}$; and 2) the transition probabilities, Π_{n+1}. As it is usual for Markov chains, we associate to the transition probabilities of this chain some random variables that we now describe. In the first place, there is the out-degree, K_{out}, of a new node. Secondly, we consider at each time $n+1$ a sequence of independent and identical distributed bivariate random vectors $\{\boldsymbol{Z}_i\}$, taking value (j, k), $j, k \in \mathbb{N}$, with probability $\Pi_{n+1}^{j,k}$, which depends on the state of the chain at time n. Based on the previous random variables, the growing network dynamics can be written as:

$$\mathcal{N}_{n+1}^{j,k} = \mathcal{N}_n^{j,k} + \Delta_n^{j,k} \quad \forall j \geq k \in \mathbb{N} \tag{5}$$

where

$$\Delta_n^{j,k} = \begin{cases} \sum_{i=1}^{K_{out}} \delta_{\boldsymbol{Z}_i=(j-1,k)} - \delta_{\boldsymbol{Z}_i=(j,k)} & \text{for } j > k \\ \delta_{K_{out}=j} - \sum_{i=1}^{K_{out}} \delta_{\boldsymbol{Z}_i=(j,j)} & \text{for } j = k. \end{cases} \tag{6}$$

The random vector $\boldsymbol{Z}_i$ indicates to which type of node the i link (of the new node) is pointing to. For example, if $\boldsymbol{Z}_1 = (3, 2)$, a new link is pointing to an existing node with 2 out-links and 1 in-link (or 3 total links). Clearly, in order to have a good representation of the growing network process, the probability law of Z_i must be equal to $\Pi_{n+1}^{j,k}$, as we impose. Equations (5) and (6) can be read in the following way: if at time $n+1$ a new node with $K_{out} = m$ out-links is aggregated, then $\mathcal{N}_{n+1}^{m,m}$ grows by one, and m components of the degree vector undergo a "shift". As the network continues to grow, the goal is to find whether there exists a limit distribution for the in-out degree. For very large values of n, given a randomly selected node, what is the probability that this one has k links, of which k_{out} are out-links, $\mathcal{P}(k, k_{out})$?

The traditional approach [5,11,20] for finding stationary probabilities is based on the Kolmogorov rate equation. Here, we present a complementary technique. The following property shows a way of computing $\mathcal{P}(k, k_{out})$ which has interest on itself.

Property. $\mathcal{P}(k, k_{out})$ is the solution of:

$$\mathcal{P}(k, k_{out}) = \langle \Delta_n^{k,k_{out}} / \Theta_n \rangle \quad \forall k \geq k_{out} \in \mathbb{N}, \tag{7}$$

where Θ_n is the event that imposes that the empirical distribution at time n is equal to the stationary distribution, i.e. $\Theta_n = \{\frac{\mathcal{N}_n^{h,i}}{\sum_{l,m\in\mathbb{N}} \mathcal{N}_n^{l,m}} = \mathcal{P}(h, i) \quad \forall h, i \in \mathbb{N}\}$. The previous property says that if the process at time n is distributed according to the stationary probability, $\mathcal{P}$, it will remain there in expectation.

Using the property mentioned above and equation (6), it is easy to see that the stationary joint deg-out distribution, $\mathcal{P}$, satisfies:

$$\begin{aligned} \mathcal{P}(k, k_{out}) &= \Pi^{k-1,k_{out}} \langle K_{out} \rangle - \Pi^{k,k_{out}} \langle K_{out} \rangle \\ \mathcal{P}(k, k) &= P(K_{out} = k) - \Pi^{k,k} \langle K_{out} \rangle \end{aligned} \tag{8}$$

for $k > k_{out} \in \mathbb{N}$, where $\langle K_{out}\rangle = \sum_{k_{out}=1}^{\infty} k_{out} P(k_{out})$. These two equations contain all the information about the limit joint in-out degree distribution, being a crucial result in this paper. It is important to note that since we have conditioned on the fact that at time n the process is distributed according to the stationary probability, the link attachment probability does not depend on time. Now, $\Pi^{k,k_{out}}$ denotes the stationary probability that a new link (from a new node) point to an existing node with $k - k_{out}$ in-degree links (or k total links) and k_{out} out-degree links. Under preferential linking on degree with attractiveness (Eq. (3)), the stationary attachment law remains:

$$\Pi^{k,k_{out}} = \frac{k + A}{\langle K\rangle + A}\mathcal{P}(k, k_{out}). \tag{9}$$

The marginal distribution of equation (9), $\pi^k = \sum_{k_{out}=1}^{k} \Pi^{k,k_{out}}$, is the stationary version of π^k_{n+1} presented in equation (3). Replacing equation (9) in equation (8), and using $\langle K\rangle = 2\langle K_{out}\rangle$ (for each new node with k_{out} out-links, the total degree increases by $2k_{out}$) we obtain:

$$\mathcal{P}(k, k_{out}) = \frac{\Psi(k + A, 3 + \delta)}{\Psi(k_{out} + A, 2 + \delta)}P(k_{out}), \tag{10}$$

where $\Psi(a, b) \equiv \frac{\Gamma[a]\Gamma[b]}{\Gamma[a+b]} = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ (Beta function), and $\delta = A/\langle K_{out}\rangle$. From equation (10), taking marginal distributions, it is trivial to obtain:

$$P(k_{in}, k_{out}) = \frac{\Psi(k_{in} + k_{out} + A, 3 + \delta)}{\Psi(k_{out} + A, 2 + \delta)}P(k_{out}) \qquad \text{(a)}$$

$$P(k) = \Psi(k + A, 3 + \delta)\sum_{k_{out}=1}^{k} \frac{P(k_{out})}{\Psi(k_{out} + A, 2 + \delta)} \qquad \text{(b)}$$

$$P(k_{in}) = \sum_{k_{out}=1}^{\infty} P(k_{out})\frac{\Psi(k_{out} + k_{in} + A, 3 + \delta)}{\Psi(k_{out} + A, 2 + \delta)}. \qquad \text{(c)}$$

$$\tag{11}$$

Equation (11) shows the joint stationary in-out degree probability, the degree distribution and the in-degree distribution. In the stationary regime (for the probability) the proportion of nodes with k_{in} in-links and k_{out} out-links (Eq. (11a)), depends on the attractiveness, and on the out-degree distribution through two quantities: $\langle K_{out}\rangle$ and $P(k_{out})$. The same happens for $P(k)$ and $P(k_{in})$. Equation (11b) gives the stationary degree probability for arbitrary out-degree distribution (see Appendix A for a simpler derivation). Note that just by replacing $P(k_{out})$ by $\delta_{k_{out}=m}$ (this means a non-random K_{out} and equal to m) we obtain the known result [5] for undirected networks. Equation (11c) constitutes one of the main results of the paper. Replacing $P(k_{out})$ by the empirical value, we can check whether the model is adequate for the network under study. Moreover, it is possible to see that a superposition principle does not hold, either for $P(k)$, $P(k_{in})$, or $P(k_{in}, k_{out})$. They cannot be written as $P(k) = \sum_{k_{out}=1}^{\infty} P(k_{out})Q_{k_{out}}(k)$, where $Q_{k_{out}}(k)$ is the stationary probability for a fixed number k_{out} of out-links. The superposition principle will be valid for the three limit distributions only when the attractiveness vanishes (preferential linking). In this way, the preferential linking generalization (the inclusion of attractiveness) introduced in [5] has the advantage of enlarging the power exponent values of the degree distribution, with the drawback of loosing a superposition principle. If we allow the appearance of new nodes with zero out-links ($P(K_{out} = 0) > 0$), then the results presented in equations (11b) and c, still hold after switching the initial index in the summation from 1 to 0 and taking $k_{out} \in \mathbb{N}_o$. In this last case, the attractiveness must be greater or equal zero (see Eq. (4)).

2.2 Descriptive statistics

Before trying to describe a real network by a model, some first checks are recommendable. One typical measure that has been extensively used is the clustering coefficient, that is a measure of how connected the neighbors of a node are. We are going to discuss much simpler descriptive measures that are useful tools for looking for the "best" model. Therefore, it is important to have analytical devices for comparing with real data in the search of a good model.

2.2.1 Covariance and conditional expectation

A measure of dependence between the in-degree and the out-degree can give an idea of which is the attachment law that better describes the empirical data. The covariance between K_{out} and K_{in}, $\mathrm{Cov}(K_{in}, K_{out}) = \langle K_{in}K_{out}\rangle - \langle K_{in}\rangle\langle K_{out}\rangle$ is an adequate statistical measure for this purpose. For example, in the case where the law of attachment is preferential linking on in-degree (Eq. (25)) this measure is obviously zero (K_{in} and K_{out} are independent). For the case studied in detail here, preferential linking on degree (Eq. (3)), it is straightforward to see that the covariance between K_{out} and K_{in} in the particular case $A = 0$, satisfies the following equation:

$$\mathrm{Cov}(K_{in}, K_{out}) = \frac{1}{2}\mathrm{Cov}(K, K_{out}) = \mathrm{Var}(K_{out}) \tag{12}$$

where $\mathrm{Var}(K_{out}) = \mathrm{Cov}(K_{out}, K_{out})$. The covariance between K_{out} and K_{in} is always positive or zero (for non random K_{out}), as it is expected for this type of attachment law. Equation (12) instead can be written in terms of the correlation, $r = \frac{\mathrm{Cov}(K_{in},K_{out})}{\sqrt{\mathrm{Var}(K_{in})\mathrm{Var}(K_{out})}}$, in the following way:

$$r = \sqrt{\frac{\mathrm{Var}(K_{out})}{\mathrm{Var}(K_{in})}}. \tag{13}$$

It is surprising that the correlation between K_{in} and K_{out} satisfy this simple relationship between the standard deviations, r is the ratio between $\sqrt{\mathrm{Var}(K_{out})}$ and $\sqrt{\mathrm{Var}(K_{in})}$.

Since the correlation coefficient is always less or equal 1, we obtain the following inequality:

$$\mathrm{Var}(K_{out}) \leq \mathrm{Var}(K_{in}). \tag{14}$$

Although it is very easy for real network to estimate the variance of the number of out and in links, and also the covariance (or correlation) between the in and out-degree, these measures are not typically reported (see Appendix B for results on the WWW network).

On the other hand, the first right term of the covariance always satisfies:

$$\langle K_{in} K_{out} \rangle = \sum_{k_{out} \in \mathbb{N}} k_{out} \langle K_{in}/K_{out} = k_{out} \rangle P(k_{out}), \tag{15}$$

where $\langle K_{in}/K_{out} = k_{out} \rangle$ is the conditional expectation of the number of in-links given that the node has k_{out} out-links. From equations (12) and (15) it is very easy to see that:

$$\langle K_{in}/K_{out} \rangle = \frac{1}{2} \langle K/K_{out} \rangle = K_{out}. \tag{16}$$

The relationship between $\langle K_{in}/K_{out} \rangle$ and K_{out} can be a second check to make before modeling. For a real network this can be done in the following way, choose all the nodes that have a number K_{out} of outgoing links, and take the mean of the number of in-links over this set of nodes. If the conditional mean is equal to K_{out} for all values of K_{out}, then this is an indication that the model can be adequate.

For non null attractiveness it is hard to obtain analytical results, nevertheless, we compute numerically $\langle K_{in}/K_{out} \rangle$ for different values of K_{out} and attractiveness. From equation (11a) and the definition of conditional expectation, it is easy to obtain:

$$\langle K_{in}/K_{out} \rangle = \sum_{j \in \mathbb{N}} j \frac{\Psi(j + K_{out} + A, 3 + \delta)}{\Psi(K_{out} + A, 2 + \delta)}. \tag{17}$$

Figure 2a shows the numerical results of $\langle K_{in}/K_{out} \rangle$ based on equation (17). For any value of the attractiveness and $\langle K_{out} \rangle$, the conditional expectation follows a linear relation with K_{out}:

$$\langle K_{in}/K_{out} \rangle = f(A, \langle K_{out} \rangle) K_{out} + g(A, \langle K_{out} \rangle). \tag{18}$$

The slope, $f(A, \langle K_{out} \rangle)$, and the intercept, $g(A, \langle K_{out} \rangle)$, of this straight line satisfies:

$$\begin{aligned} &\lim_{A \to \infty} f(A, \langle K_{out} \rangle) = 0 \\ &\lim_{A \to \infty} g(A, \langle K_{out} \rangle) = \langle K_{out} \rangle, \end{aligned} \tag{19}$$

as it is shown in Figure 2b and 2c. For positive values of attractiveness the slope is smaller than one, going to zero as the attractiveness goes to infinity. In the case $A \to \infty$, K_{in} and K_{out} are independent (always with the same expectation). Finally, for negative values of A the slope is greater than one. Studying the empirical relationship between $\langle K_{in}/K_{out} \rangle$ and K_{out} can give some insight on the model. Moreover, if this relationship is linear, from Figure 2b and 2c, it is possible to have a first estimation of the attractiveness. In Appendix B we show the statistical measures presented here for the WWW network.

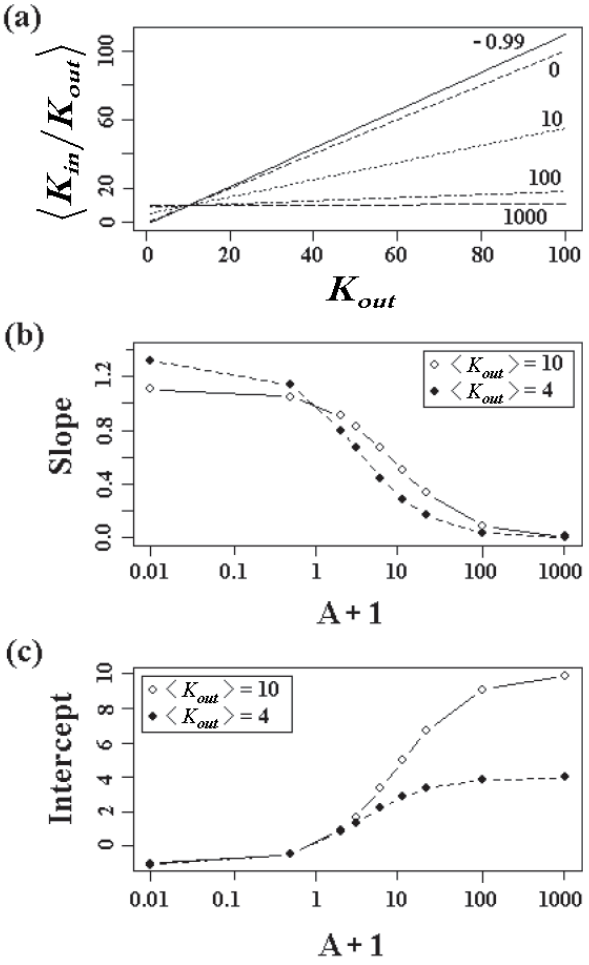

Fig. 2. (a) Conditional expectation of in-degree given the out-degree. Each straight line correspond to a different value of attractiveness (specified in the graph). (b) Slope and (c) Intercept of the type of straight lines shown in (a) as a function of the attractiveness for two different values of $\langle K_{out} \rangle$.

It is important to note that equations (12) (which includes (13), (14), and (18) (which include (16)) holds for any out-degree distribution ($P(k_{out})$). These results do not depend on the details (shape) of the out-degree distribution. Nevertheless, there exist some measures that do not share this nice property. For example, the conditional number of out-links given the number of in-links, $\langle K_{out}/K_{in} \rangle$, depends explicitly on $P(k_{out})$, as can be seen in the following equation:

$$\langle K_{out}/K_{in} \rangle = \frac{\sum\limits_{k_{out}=1}^{\infty} k_{out} \frac{\Psi(K_{in}+k_{out}+A,3+\delta)}{\Psi(k_{out}+A,2+\delta)} P(k_{out})}{\sum\limits_{h_{out}=1}^{\infty} \frac{\Psi(K_{in}+h_{out}+A,3+\delta)}{\Psi(h_{out}+A,2+\delta)} P(h_{out})}. \tag{20}$$

Next, we present another measure useful for model selection.

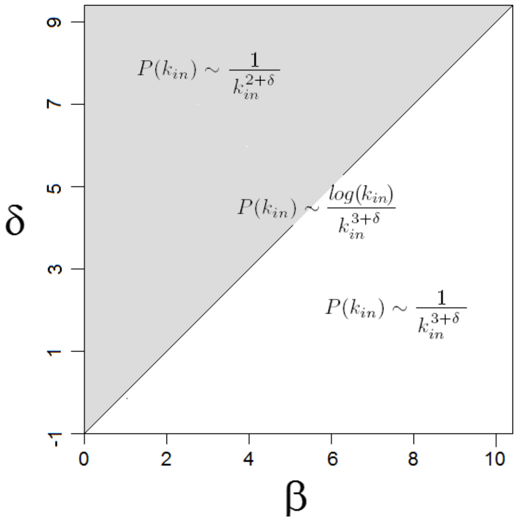

Fig. 3. Stationary in-degree probability tail under preferential linking with attractiveness, for an out-degree with $P(k_{out}) \sim \frac{1}{k_{out}^{2+\beta}}$, as a function of $\delta = \frac{A}{\langle K_{out} \rangle}$ and β. The horizontal axis corresponds to preferential linking ($A = 0$). In the separatrix curve, $\delta = \beta - 1$, $P(k_{in}) \sim \frac{log(k_{in})}{k_{in}^{3+\delta}} = \frac{log(k_{in})}{k_{in}^{2+\beta}}$.

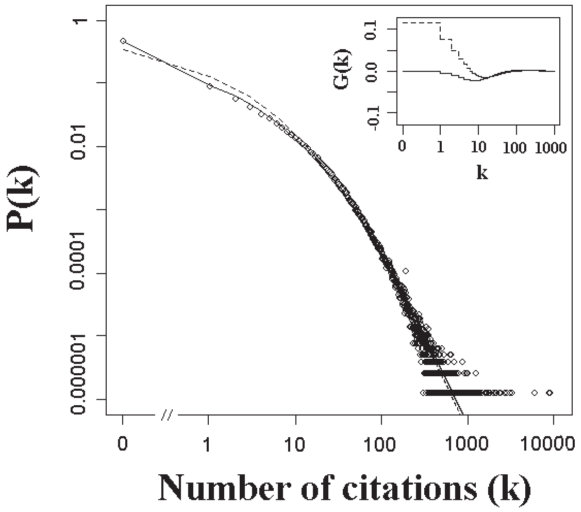

Fig. 4. Citation distribution for all papers published in 1981 (from the ISI) cited between 1981 and 1997. The theoretical citation (in-degree) curves are calculated by equation (11c) assuming that $A = 0$, and the out-degree distribution is geometric, $P(k_{out}) = p(1-p)^{k_{out}}$ for $k \in \mathbb{N}_o$. The dashed line corresponds to $p = 0.104$ ($T = 0.115$). The solid line corresponds to $P(k_{out}) = 0.7622781p(1-p)^{k_{out}}$ for $k \in \mathbb{N}$ and $P(K_{out} = 0) = 0.3$, with $p = 0.0817$ ($T = 0.023$). Inset: difference between the empirical cumulative distribution and the theoretical cumulative distribution. Data from [16].

2.2.2 Relationship between the distribution tails

Now, we study the relationship between the tails of the in-degree and the out-degree distributions. In the case $A = 0$, if the out-degree distribution has finite expectation ($\langle K_{out} \rangle < \infty$) and a scale invariant tail, $P(k_{out}) \sim k_{out}^{-(2+\beta)}$, it is not difficult (from Eq. (11b)) to see that the limit degree distribution and the in-degree distribution have the following tail behavior:

$$P(k_{in}) \sim P(K = k_{in}) \sim \begin{cases} k_{in}^{-(2+\beta)} & 0 < \beta < 1 \\ log(k_{in})k_{in}^{-3} & \beta = 1 \\ k_{in}^{-3} \quad \beta > 1. \end{cases} \tag{21}$$

Equation (21) constitutes our second main result: if the out-degree distribution has finite variance and a scale invariant tail, $P(k_{out}) \sim k_{out}^{-(2+\beta)}$, then the limit in-degree distribution has also a scale invariant tail, $P(k_{in}) \sim k_{in}^{-\alpha}$. Moreover, for $0 < \beta < 1$, α is equal to the out-degree exponent. This last result can explain why in so many real networks the in and out power exponents are so similar, taking values in a range from 2 to 3. In the case $\beta > 1$, $\alpha = 3$, regardless of the value of β. For the frontier case (finite/infinite variance) of $\beta = 1$, the limit distribution decays at a slower rate than k_{in}^{-3}. Precisely, it decays as $P(k_{in}) \sim \log(k_{in})k_{in}^{-3}$. In the general case of preferential linking with attractiveness for $P(k_{out}) \sim k_{out}^{-(2+\beta)}$, the regimes are similar to the non-attractiveness case. In this case, there is a separatrix curve between the in-degree behaviors, as it is shown in Figure 3. The behavior is regulated by $\delta \equiv A/\langle K_{out} \rangle$ and β. For $\delta > 1+\beta$, the (in) degree distribution has exactly the same tail as the out-degree ($P(K_{out} = k) \sim P(K_{in} = k) \sim P(K = k) \sim k^{-(2+\beta)}$), even for large β. For $\delta < 1+\beta$, $P(k_{in})$ behaves as $k_{in}^{-(3+\delta)}$. Finally on the separatrix curve, $\delta = 1+\beta$, the asymptotic behavior goes as $\log(k_{in})k_{in}^{-(3+\delta)}$. Note that δ can not be smaller than -1, since $\langle K_{out} \rangle$ must be (see Eq. (4)) greater than –A.

For out-degree distributions with exponential tails, as a geometric, Poisson, or finite range distributions, the in-degree distribution satisfies that $P(k_{in}) \sim k_{in}^{-(3+\delta)}$, even for negatives values of δ. In [12] they show that the PRL citation network has an out-degree distribution with exponential decay, and an in-degree distribution with a power law tail with α near 3, just as described before for the null attractiveness case. We remark the following: a) if the model is adequate for describing a real growing network, and this network has an out-degree distribution with exponential tail, and a scale invariant in-degree distribution with a power between 2 and 3, then attractiveness parameter must be negative; and b) if the empirical in-degree distribution has a scale invariant tail with a power less than 2, then the model presented here is not adequate for describing this network. Keeping in mind the last point, the new estimations [13] of the in-degree power exponent of the WWW network, would rule out the model for describing this particular network.

2.3 Application: scientific publications network

The scientific publications network has two advantages that define it as the most “pure”: 1) extremely few double arrows; and 2) all the variability in the number of outlinks is "intrinsic". These two features guarantee that our model (see Fig. 1) is adequate for describing the scientific network. Nevertheless, it is not clear which is the attachment law (π) such that we can obtain a good mimic of the growing network process.

Figure 4 shows the citation distribution for all (1981) scientific publications published (from the ISI dataset) cited between 1981 and 1997 (see [6]). Clearly, this distribution represent the in-degree one of a growing directed network. We suppose this distribution correspond to the stationary case[1].

Unfortunately the out-degree distribu tion ($P(k_{out})$), the number of references that has a randomly selected paper, has not been reported, making impossible a plug-in approach (see Eq. (11c)) to test the growing model. Nevertheless, we take the following strategy: we suppose a geometric out-degree distribution $P(k_{out}) = p(1-p)^{k_{out}}$ with $k \in \mathbb{N}_o$, a preferential linking on degree attachment law (Eq. (3) with $A = 0$), and finally we estimate p. Probably the empirical out-degree distribution ($P(k_{out})$) does not fall in any family of parametric distributions. However, a well estimated in-degree distribution will be a positive result, since the in-degree distribution is obtained as a result of a theoretical computation based on the out-degree distribution. In order to estimate p, we first compute the average number of citations in the ISI network ($\langle citations \rangle = 8.573$) and impose the condition:

$$\langle citations \rangle = \langle references \rangle. \tag{22}$$

With this condition ($\langle K_{in} \rangle = \langle K_{out} \rangle$) we obtain that $p = 1/(9.573)$. The dashed line in Figure 4 corresponds to this case. If we estimate separately the case $k = 0$, and assume that the out-degree distribution is such that $P(K_{out} = 0) = a$, and $P(k_{out}) = cp(1-p)^{k_{out}}$ for $k \in \mathbb{N}$ with $c = (1-a)/(1-p)$, we obtain $p = (1-a)/8.573$ after taking the mean value condition (Eq. (22)). Curiously, for $a = 0.3$ ($p = 0.0817$) the theoretical in-degree probability (solid line) is extremely similar to the empirical one in all the range of the distribution. This can not be achieved with an oversimplified model where $P(k_{out}) = \delta_{k_{out}=m}$. But, this is not the only $P(k_{out})$ that fits perfectly well, hence we do not assert that the estimated $P(k_{out})$ must be similar to the real citation distribution. Moreover, the estimated $P(k_{out})$ does not seem very adequate, since under this probability distribution 30% of all scientific publications do not contain any reference (yet, note that in [10] it was reported that 10% of all publications do not contain any reference).

In order to have a better notion of the goodness of fitness we compute the Kolmogorov statistic,

$$T = \underset{k \in \mathbb{N}_o}{max} |G(k)| = \underset{k \in \mathbb{N}_o}{max} |F_{\widehat{P}_{in}}(k) - F_{P_{in}^{theo}}(k)|, \tag{23}$$

where $F_P(k)$ is the cumulative distribution,

$$F_P(k) = \sum_{j=0}^{k} P(j). \tag{24}$$

P_{in}^{theo} correspond to the theoretical in-degree distribution showed in equation (11)(c) assuming a particular $P(k_{out})$, and $\widehat{P}_{in}$ correspond to the empirical citation distribution. One advantage of the T statistic (Eq. (23)) is that it is now possible to test whether the model (including the attachment law) is adequate for describing the real network. In this growing network application, the null hypothesis is H_o: the real growing network has an underlying link attachment law that is preferential on degree. For the simplest case where T compares an empirical distribution with a theoretical one, but without estimating parameters, the null hypothesis will be rejected (at a 0.05 level of significance) only if $T > 0.0015$. In the case shown with solid line $T = 0.023$, and for the case where $P(k_{out})$ is geometric (dashed line) $T = 0.115$. Clearly, T is a good measure for ranking models (or model selection). The inset of Figure 4 shows the function $G(k)$ for both out-degree distributions proposed. For the geometric (dashed line) case the maximum distance between the cumulative distributions (see Eq. (23)) occurs at $k = 0$, and for the other case (solid line) at $k = 10$.

As we mentioned at the beginning of this section, the model is adequate for the scientific publication network, but the attachment law is completely unknown. We have proposed one, preferential linking on degree, but we do not have the possibility to corroborate it. This is one of the reasons why we are going to study the model under different attachment laws. The only weak argument in favor of the law given by equation (3), is that review papers, that have a huge number of references, are typically highly cited [17] compared with regular articles that have a small number of references. In this way, the correlation between K_{in} and K_{out} will be positive, which is a virtue of the law defined in equation (3).

2.4 Different attachment laws

Clearly, it may happen that for a real network the informal checks (covariance, variance and conditional expectation) discussed before might be inconsistent with the observables of the model. In this case, three things may be happening: 1) the link attachment law is not adequate; 2) the model is not correct; or 3) both before. The first point is related to the mechanism of linking: preferential, uniform, non linear preferential, or may have some age dependency as described in [18,19]. The second point correspond to the growing mechanism, that can be seen as the core of the model. For example, updating of nodes, or a very high proportion of double links can be present, that are not considered in the model. In this section we discuss only the alternative where the attachment law is different from the one proposed in equation (3) (preferential linking on degree), but the core of the model remains true.

[1] If the scientific network has arrived (in 1997) to a proportion of papers with k citations that do not change with time (stationary), then the articles published in a particular year (1981) are a sample of this distribution

2.4.1 Preferential linking on in-degree

In [5] the authors studied the B-A model when the attachment law depends on the degree and on the attractiveness. The proposed law can be expressed in the following way [2]:

$$\pi_{in}^{k_{in}} = \frac{(k_{in}+A)N_{in}^{k_{in}}}{\sum\limits_{h_{in}\in\mathbb{N}_o}(h_{in}+A)N_{in}^{h_{in}}}, \tag{25}$$

where $N_{in}^{k_{in}}$ is the number of nodes with in-degree equal k_{in}, and now the attractiveness (A) is greater or equal zero. In principle, this can be a good attachment law for the scientific publications network. The joint attachment law in this case is given by:

$$\varPi^{k,k_{out}} = \frac{k-k_{out}+A}{\langle K_{out}\rangle + A}\mathcal{P}(k,k_{out}), \tag{26}$$

where we have used that $\langle K_{in}\rangle = \langle K_{out}\rangle$. Replacing equation (26) in equation (8), it is very easy to compute the stationary probabilities:

$$\begin{aligned}
P(k_{in}) &= \frac{\varPsi(k_{in}+A, 2+\delta)}{\varPsi(A, 1+\delta)}\\
P(k) &= \frac{1}{\varPsi(A,1+\delta)}\sum_{k_{out}=0}^{k} P(k_{out})\varPsi(k-k_{out}+A, 2+\delta)\\
P(k_{in},k_{out}) &= \mathcal{P}(k_{in}+k_{out},k_{out}) = P(k_{in})P(k_{out})
\end{aligned} \tag{27}$$

where $k, j \in \mathbb{N}_o$. This case is specially easy to solve because, for a randomly selected node, the number of out-links (K_{out}) and the number of in-links (K_{in}) are independent random variables ($P(k_{in},k_{out}) = P(k_{in})P(k_{out})$). This mean:

$$\begin{aligned}
&r = 0 && (a)\\
&\langle K_{in}/K_{out}\rangle = \langle K_{out}/K_{in}\rangle = \langle K_{out}\rangle. && (b)
\end{aligned} \tag{28}$$

One big difference between the previous attachment law (Eq. (3)) and this one (Eq. (25)) is that $P_{in}(k)$ depends only on the mean number of out-links ($\langle K_{out}\rangle$) by δ ($A/\langle K_{out}\rangle$), and not on the shape of the out-degree distribution. For $A > 0$ and $k_{in} \gg 1$, $P(k_{in})$ behaves as $k_{in}^{-(2+\delta)}$ no matter which is $P(k_{out})$ (only depends on $\langle K_{out}\rangle$). Therefore, under the attachment law given by equation (25), the tail of the out-degree distribution does not give any information about the tail of in-degree distribution, contrary to what happens for the law of equation (3). In addition, for this new attachment law the correlation between K_{in} and K_{out} is zero (Eq. (28a)), and the conditional expectation of K_{in} (K_{out}) given K_{out} (K_{in}) does not depend on K_{out} (K_{in}) (Eq. (28b)).

Note that $\pi_{in}^{k_{in}}$ is well defined only for positive or zero values of attractiveness. But, only strictly positive values of A are interesting, since for $A = 0$ we get that the stationary probability is $P(k_{in}) = \delta_{k_{in}=0}$. This last result is easy is to understand: new nodes appear but they can not be pointed by other nodes ($A = 0$), and in this way the network will be formed by almost all nodes with zero in-links and only a few (given by the initial condition of the network) with many in-links. Clearly, in the limit $n \to \infty$ the proportion of nodes with k_{in} in-links goes to a delta function ($\delta_{k_{in}=0}$).

2.4.2 Uniform attachment law

It is thus clear that even when preferential linking is an accepted mechanism of link attachment, it is necessary to study [20,21] alternative types. For the uniform attachment law on degree:

$$\begin{aligned}
\pi_{unif}^{k} &= \frac{N^k}{\sum\limits_{j\in\mathbb{N}} N^j}\\
\varPi^{k,k_{out}} &= \mathcal{P}(k,k_{out})
\end{aligned} \tag{29}$$

by means of the same technology (replacing $\varPi^{k,k_{out}}$ in Eq. (8)) we obtain:

$$\begin{aligned}
P(k) &= \frac{1}{1+\langle K_{out}\rangle}\sum_{k_{out}=0}^{k} P(k_{out})\left(\frac{\langle K_{out}\rangle}{1+\langle K_{out}\rangle}\right)^{k-k_{out}}\\
P(k_{in}) &= \frac{1}{1+\langle K_{out}\rangle}\left(\frac{\langle K_{out}\rangle}{1+\langle K_{out}\rangle}\right)^{k_{in}}.
\end{aligned} \tag{30}$$

Note that, $P(k_{in})$ depends only on $\langle K_{out}\rangle$ (and not on $P(k_{out})$), and decays exponentially fast. For an out-degree with $P(k_{out}) \sim k_{out}^{-(2+\beta)}$, $P(k)$ behaves as $k^{-(2+\beta)}f(k)^{-1}$, where $f(k)$ is an increasing function of k that grows more slowly than $log(k)$. It is important to remark that for empirical (finite) networks, the $f(k)^{-1}$ term will be very difficult to discriminate ($f(k)$ grows at a rate slower than $log(log(k))$). This behavior may be hard to "separate" from $P(k) \sim k^{-(2+\beta)}$, but the in-degree distribution will sort out any possible confusion about the link attachment law.

2.5 Implementation of the model

Being rigorous, the model as it was presented in Section 2 is not well defined. Yet, as we discuss in this section, this is not a serious problem (all the results presented before hold). The difficulty is that $P(k_{out})$ is any probability distribution. In particular, it includes the ones that take infinitely values (such as geometric, or any one with exponential or power law tails). The problem can be stated as follows: if a new node, for example has 1000 links and the network has 100 nodes, ¿what do we must do with the remaining 900 links?

[2] For a fixed number of out-links equal m (B-A model), $\frac{(A+k)N_{in}^{k}}{\sum_{j\in\mathbb{N}_o}(A+j)N_{in}^{j}} = \frac{(A+k)N^{k+m}}{\sum_{j\in\mathbb{N}_o}(A+j)N^{j+m}} = \frac{(\tilde{A}+k+m)N^{k+m}}{\sum_{j\geq m}(\tilde{A}+j+m)N^{j+m}}$, where $\widetilde{A} = A - m$. The attachment law of equation (25) is equivalent to the one of equation (2) replacing A by $A - m$.

We describe below the correct form of the model (that can be implemented):

(1) initially the network consists of n nodes connected in a given arbitrary way;
(2) at each time step starting from $n+1$, say time step r, a node with $\widetilde{K}^r_{out}$ outgoing-edges appears. $\widetilde{K}^r_{out}$ is a random variable with law $Q^r(k_{out})$ ($Q^r(k_{out}) \equiv P(\widetilde{K}^r_{out} = k_{out})$, and $\sum_{j=1}^{r} P(\widetilde{K}^r_{out} = j) = 1$);
(3) each new directed edge points out to an existing node with some probability law π_r (uniform, preferential linking, etc.).

The distribution of the number of out-links from a new node at time r (the networks has $r-1$ nodes) is defined by the following equation:

$$Q^r(k_{out}) = P(K_{out} = k_{out}/K_{out} < r). \tag{31}$$

$Q^r(k_{out})$ is the conditional distribution of K_{out} given $K_{out} \leq r-1$. From definition 31 is very easy to see that $Q^r(k_{out})$ converge to $P(k_{out})$,

$$\lim_{r\to\infty} Q^r(k_{out}) = P(k_{out}), \tag{32}$$

as the network grows, where $P(k_{out})$ is the distribution defined a priori (see Sect. 2). From this last convergence we can see that the model with this correction (we have only changed $P(k_{out})$ by $Q^r(k_{out})$) has exactly the same asymptotic behavior that was obtained for the model presented in Section 2. Therefore, all the results presented in this paper also hold for the corrected model. The general conclusion would be: "small effects disappear at ∞". See, for instance Section 2.4.1 where we discuss why does $P(k_{in})$ converge to $\delta_{k_{in}=0}$ for $A = 0$.

To illustrate, we present simulations of the corrected model. We start with a network of only 3 nodes where each of them has an out-degree and in-degree equal to one (triangle configuration), and let the network grows, under a preferential linking on degree with null attractiveness (Eq. (3)) attachment law, up till it reaches 100 000 nodes. In each temporal step a new node appears, which has a number of out-links, $\widetilde{K}^r_{out}$ (random variable), that depends on the number of nodes (r) in the network. Figure 5 shows the in-degree and out-degree distribution (points) obtained from the simulations, for different conditional probability laws ($Q^r(k_{out})$) for $\widetilde{K}^r_{out}$. For example, in Figure 5a we show the empirical (in)out-degree distribution for the case where $\widetilde{K}^r_{out}$ has a probability law $P(\widetilde{K}^r_{out} = k_{out}) = ck_{out}^{-2.1}$ (for $k_{out} \in 1, 2, ..., r$), where c is the normalization constant. The solid lines in the out-degree distributions correspond to the limit of $Q^r(k_{out})$, that is called $P(k_{out})$ (in the previous example $P(k_{out}) = 0.64093k_{out}^{-2.1}$ for $k_{out} \in \mathbb{N}$). The solid lines in the in-degree distributions correspond to the stationary (in-degree) distribution obtained in Section 2.1 (Eq. (11c) with $A = 0$ and $P(k_{out})$ the proposed law) for the "incorrect" model. Also, in each graph we present a guide reference for the power law tail behavior (dashed lines). This figure shows two important (now, not surprising) facts: 1) in order to find the stationary (in) degree distribution for the corrected model, it is enough to study the stationary distributions of the "incorrect" model; and 2) the simulations confirm the relationship, reported in Section 2.2.2, between the tails of the in-degree and out-degree distributions.

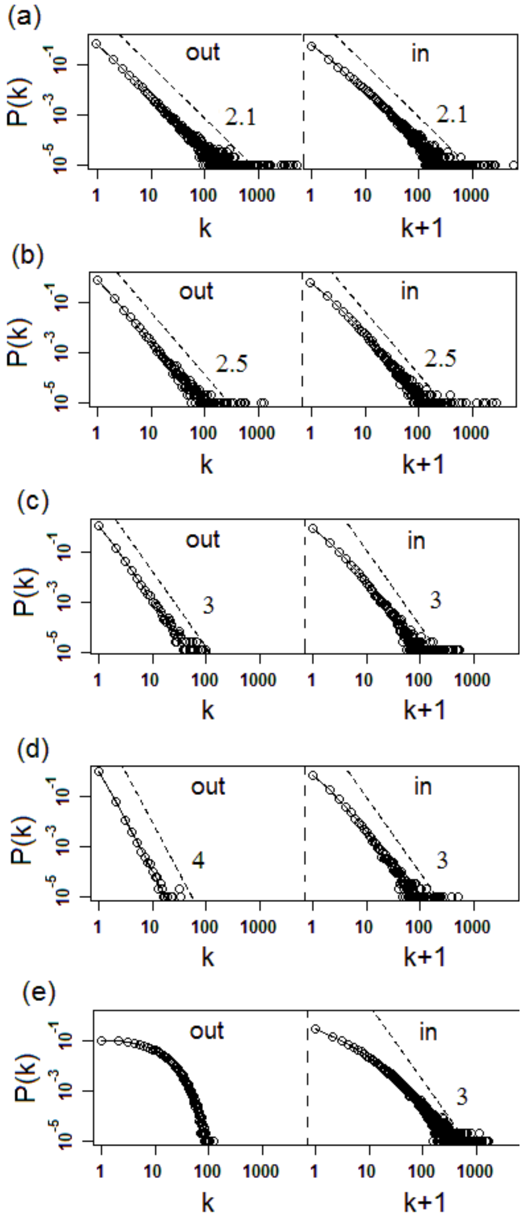

Fig. 5. In-degree and out-degree distribution (points) from simulations of the corrected model, for the cases where $Q^r(k_{out})$ is: (a) $ck_{out}^{-2.1}$, (b) $ck_{out}^{-2.5}$, (c) ck_{out}^{-3}, (d) ck_{out}^{-4}, and (e) $0.1(0.9)^{k_{out}}$, where c is the normalization constant (it depends on r). The solid lines correspond to the stationary distribution given by equation (11c) with $A = 0$ ($P(k_{in}) = \sum_{k_{out}=1}^{\infty} P(k_{out}) \frac{\Psi(k_{out}+k_{in},3)}{\Psi(k_{out},2)}$). The dashed lines are plotted as a guide reference, and the number below indicates the slope of the straight line.

3 Summary and discussion

For the model presented here, we showed a simple way to compute the stationary probabilities. This model was constructed in order to take into account the main features of real directed growing networks with the property that almost all the variability in the number of out-links is "intrinsic" (see Sect. 2). From the stationary Property, we showed how to compute the stationary joint in-out degree distribution for an arbitrary out degree distribution, and arbitrary link attachment law (π). We studied three different π's, paying special attention to the preferential linking on degree with attractiveness mechanism ($\pi^k = \frac{(A+k)N^k}{\sum\limits_{j\in\mathbb{N}}(A+j)N^j}$). Once obtained the joint probability, we compute:

(1) $P(k_{in})$ as a function of $P(k_{out})$;
(2) The correlation between K_{in} and K_{out};
(3) The conditional expectation of $K_{in}(K_{out})$ given K_{out} (K_{in}).

From $P(k_{in})$ we studied the relationship between the distribution tails, giving a possible explanation for the in/out degree tail relationship reported for many real networks. The statistical measures mentioned in (2) and (3) were studied for the WWW network, obtaining good agreement with some of the analytical results presented in this paper. Nevertheless, we cannot say that the model is appropriate to describe this network (an important part of the variability would be not "intrinsic").

Finally, we showed an application to the scientific publications network. In this network:

(a) New publications continuously appear (growing network) and do not disappear.
(b) The structure is rigid. Published papers cannot change their references, only new papers can change the number of citations of already published works.
(c) The publication that is forthcoming has a non predictable number of references, K_{out} (random variable)
(d) Even knowing K_{out}, the cited papers by the forthcoming publication are unpredictable (there is a law of attachment, π).

The model we proposed considers the four points mentioned above. The main difference with other models, is that the number of out-links (references) of a new node (paper) is treated now as a random variable. Therefore, if the distribution of the number of references ($P(k_{out})$) is known, an important part ((a),(b) and (c)) of the scientific network will be well described by the model. But, the distribution of the number of references of the forthcoming publication (out-degree distribution) has not been reported. In addition, the attachment law ((d)) of the scientific publication network is completely unknown, and difficult to estimate it. Thus, we proposed a simple out degree distribution (geometric) and an attachment law of preferential linking on degree (we also consider preferential linking on in-degree and uniform attachment). With these two assumptions, we found a very good fit. This application also served to discuss how to compare various models. In this matter, we proposed a measure (Eq. (23)) frequently used in statistics to compare two distributions.

From a modeling point of view, we see our results as a further step from which more complex models may be built in order to be closer to reality. The model can be seen as the skeleton to construct more sophisticated models. For example, it does not seem difficult to incorporate in the model double links (a mixed out-links distribution) in order to be closer to the metabolic network, or some updates in the nodes to mimic the WWW network. Other important issue to explore is what happens when $P(k_{out})$ depends on time in a simple parametric way. This last point is related with accelerating networks [22].

We thank A. Calabrese, A. Cuevas, M. O'Connell, and G. Solovey for critical reading of the manuscript, I. Armendáriz, and P. Ferrari for useful discussions, and A.L. Barabási and S. Redner for their generosity in sharing network data. Comments of anonymous referees are most appreciated.

Appendix A: A closed equation for $P(k)$

If we were only interested on the stationary degree distribution ($P(k)$), the computation is much easier than the one presented in Section 2.1, since there is a closed equation for $P(k)$. The growing network dynamics is given by:

$$\begin{aligned} N^k_{n+1} &= N^k_n + \Delta^k_n && (a) \\ \Delta^k_n &= \delta_{K_{out}=k} + \sum_{i=1}^{K_{out}} \delta_{Y_i=k-1} - \delta_{Y_i=k} && (b) \end{aligned} \qquad \text{(A.1)}$$

where $\{Y_i\}_{1\leq k\leq n}$ is a sequence of independent and identical distributed random variables, taking value k ($k\in\mathbb{N}$) with probability π^k_{n+1}.

Property. $\boldsymbol{P} \equiv (P(1), P(2), \ldots, P(k), \ldots)$ is the solution of:

$$\left\langle \Delta^k_n / \frac{\boldsymbol{N}_n}{\sum\limits_{k\in\mathbb{N}} N^k_n} = \boldsymbol{P} \right\rangle = P(k) \qquad \forall k\in\mathbb{N}. \qquad \text{(A.2)}$$

Replacing Δ^k_n by equation (A.1b) in equation (A.2), we get:

$$\left\langle \delta_{K_{out}=k} + \sum_{i=1}^{K_{out}} \delta_{Y_i=k-1} - \delta_{Y_i=k} / \frac{\boldsymbol{N}_n}{\sum\limits_{k\in\mathbb{N}} N^k_n} = \boldsymbol{P} \right\rangle = P(k). \qquad \text{(A.3)}$$

From this last equation it is trivial to obtain that the stationary degree probability satisfies:

$$P(k) = P(K_{out}=k) + (\pi^{k-1} - \pi^k)\langle K_{out}\rangle \qquad \text{(A.4)}$$

where π^k is the stationary probability that a new link attaches to a node with degree k. Under preferential linking on degree with attractiveness, the stationary attachment

Table B.1. Descriptive statistical measures for 4 WWW networks. Data from [13].

	$\mathrm{Cov}(K_{in}, K_{out})$	$\mathrm{Var}(K_{out})$	$\mathrm{Var}(K_{in})$
WBGC01	155.682	171.61	40080.04
WGUK02	524.244	750.76	20534.89
WBGC03	348.486	870.25	54980742
WGIT04	3478.75	4502.41	776866

Table B.2. Correlation (r) and R for 4 WWW networks. Data computed from Table B.1.

	r	R
WBGC01	0.0594	0.0654
WGUK02	0.1335	0.1912
WBGC03	0.0016	0.004
WGIT04	0.0588	0.0761

law, π^k, remains equal to $\frac{(k+A)P(k)}{\langle K\rangle + A}$. Replacing π^k in equation (A.4), and using $\langle K\rangle = 2\langle K_{out}\rangle$, it is easy to conclude that the limit degree distribution ($P(k)$) is given by equation (A.5).

$$P(k) = \Psi(k + A, 3 + \delta) \sum_{k_{out}=1}^{k} \frac{P(k_{out})}{\Psi(k_{out} + A, 2 + \delta)}. \qquad \text{(A.5)}$$

Appendix B: WWW network

As we have mentioned in the Section 2.2.1, it is difficult to find articles on networks that report the simple descriptive measures (covariance, variance and conditional expectation) for nodes discussed here. However, a detailed statistical analysis of the topological properties of four different WWW networks have been reported recently [13]. In [13] the covariance and the variance of the number of out-going links (K_{out}) and in-going links (K_{in}) were reported, which we give in Table B.1. The first thing that can be noted is that for all the domains studied $\mathrm{Var}(K_{out}) < \mathrm{Var}(K_{in})$, consistent with equation (14). Moreover, $\mathrm{Cov}(K_{in}, K_{out})$ and $\mathrm{Var}(K_{out})$ have similar values (consistent with Eq. (12)), the relative differences seems large only for WBGC03. In order to compare in a better way these last two quantities, Table B.2 shows r and $R := \sqrt{\frac{\mathrm{Var}(K_{out})}{\mathrm{Var}(K_{in})}}$ for the same data. We can see that WBGC01 and WGIT04 have very similar values of r and R (see Eq. (13)). In order to study the relationship between $\langle K_{in}/K_{out}\rangle$ and K_{out} is necessary to have the complete data. At this point, we analyze the WWW data obtained from [14] presented in [15]. We built up a database with the information of the number of out-links and in-links ((K_{out}, K_{in})) for each of the 325 729 nodes. In order to have a good estimation of the conditional expectation, we first restrict the study to the values of K_{out} such that there exist at least 500 nodes. Figure B.1a shows the relationship between K_{out} and the conditional mean of K_{in} ($\langle K_{in}/K_{out}\rangle$) given K_{out}. Interestingly, there is a strong relationship between both. For values of the K_{out} smaller than 20 there is a clear linear relationship between them. A robust regression (least median of squares) estimation between $\langle K_{in}/K_{out}\rangle$ and K_{out} gives a slope of 0.523 and an intercept of 1.739. In the case K_{out} is greater than 20 it seems that $\langle K_{in}/K_{out}\rangle$ grows faster than linear, but it is not clear if this effect is real (based in Fig. B.1b). The graph presented in Figure B.1b is similar to the one in (a),

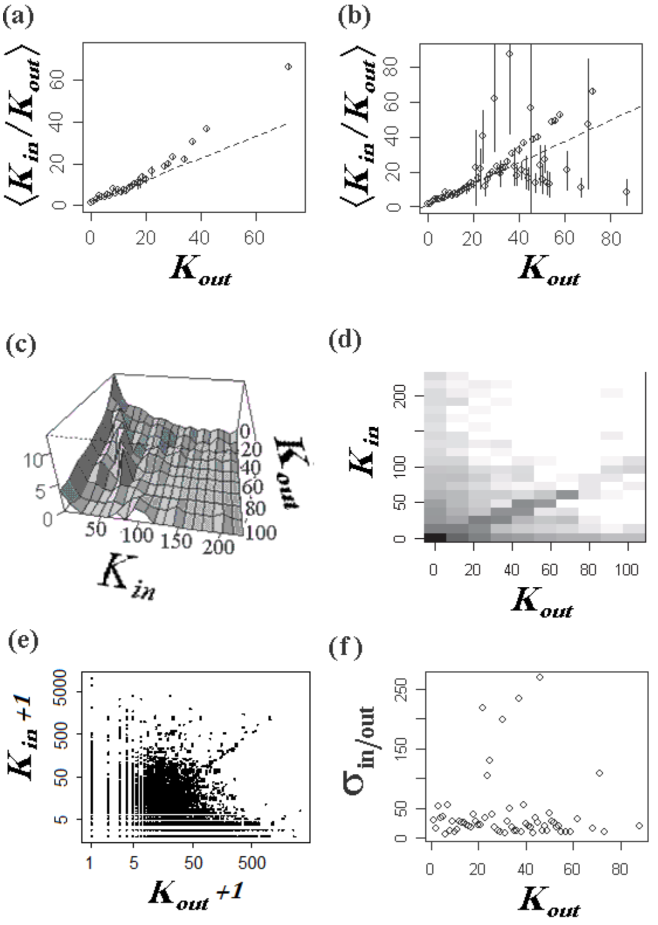

Fig. B.1. Conditional mean of K_{in} given K_{out}, when for each value of K_{out} there exist at least: (a) 500, and (b) 30 nodes. Data presented as a confidence interval of 95%. (c) and (d) Different representations of the joint in-out density of the links in a node. (e) Scatter plot of K_{in} as a function of K_{out}. (f) Conditional standard deviation of K_{in} given K_{out}, $\sigma_{in/out}$. Data from [14].

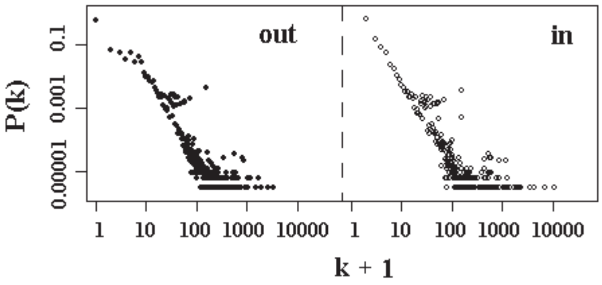

Fig. B.2. $P(K_{out} = k + 1)$ and $P(K_{in} = k + 1)$ as a function of $k + 1$. This graph was presented in [15].

but now we study the values of K_{out} such that there exist at least 30 nodes. A plot of two different representations of the joint in-out distribution is given in Figures B.1c and B.1d, to have an idea of the shape of the joint law, while (e) shows a scatter plot on a larger grid. Besides, the in-degree variance ($\mathrm{Var}(K_{in}) = 1346.85$) is greater than the out-degree one ($\mathrm{Var}(K_{out}) = 461.25$), consistent with equation (14). Figure B.1f shows the conditional standard deviation of K_{in} given K_{out}, $\sigma_{in/out} = \sqrt{\mathrm{Var}(K_{in}/K_{out})}$. Unlike the conditional expectation, the conditional variance does not seem to have any relationship with K_{out}.

In [15] the authors showed the empirical out-degree ($P(k_{out})$) and in-degree ($P(k_{in})$) distributions (see Fig. B.2), and reported a power exponent of 2.45 for the out-degree distribution and a value of 2.1 for the in-degree[3]. This is the first empirical evidence that the model presented here cannot describe well the WWW network. The model has the characteristic that the power law exponents (in-out) are equal. Finally, the second evidence that contradict the model is the fact that in this network r and R are not similar, $r = 0.2244$ and $R = 0.5852$.

References

1. A.L. Barabási, R. Albert, Science **286**, 509 (1999)
2. R. Albert, A.L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
3. S.N. Dorogovtsev, J.F.F. Mendes, Adv. Phys. **51**, 1079 (2002)
4. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabási, Nature **407**, 651 (2000)
5. S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000)
6. S. Redner, Eur. Phys. J. B **4**, 131 (1998)
7. P.L. Krapivsky, G.J. Rodgers, S. Redner, Phys. Rev. Lett. **86**, 5401 (2001)
8. B. Tadić, Physica A **293**, 273 (2001)
9. M.E.J. Newman, SIAM Review **45**, 167 (2003)
10. D.J. Price, Science **149**, 510 (1965)
11. P.L. Krapivsky, S. Redner, Lect. Notes in Physics **625**, 1616 (2003)
12. R. Lambiotte, `http://www.lambiotte.be/talks/vienna2006.pdf` (2006)
13. M.A. Serrano, A. Maguitman, M. Boguña, S. Fortunato, A. Vespignani, *ACM Trans. Web*, **1**, No.2. Article 10 (2007)
14. `http://www.nd.edu/$\sim$networks/resources.htm`
15. R. Albert, H. Jeong, A.L. Barabási, Nature **401**, 130 (1999)
16. `http://physics.bu.edu/$\sim$redner/projects/citation/isi.html`
17. J.M. Soler, J. Informetrics **1**, 123 (2007)
18. S.N. Dorogovtsev, J.F.F. Mendes, Phys. Rev. E, **62**, 1842 (2000)
19. K.B. Hajra, P. Sen, Physica A, **346**, 44 (2005)
20. P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett., **85**, 4629 (2000)
21. P.L. Krapivsky, S. Redner, Phys. Rev. E, **63**, 66123 (2001)
22. S.N. Dorogovtsev, J.F.F. Mendes, *Handbook of Graphs and Networks: From the Genome to the Internet* (Wiley-VCH, Berlin, 2002), p. 318

[3] Our estimations of the power law exponents have some differences from the ones in [15], but the difference between the in and out exponents is still appreciable.