Mixing properties of growing networks and Simpson's paradox

Andrea Capocci[1] and Francesca Colaiori[2]
[1]*Centro Studi e Ricerche E. Fermi, Compendio Viminale, Roma, Italy*
[2]*Dipartimento di Fisica, Università di Roma "La Sapienza" and SMC, INFM UdR Roma1 Piazzale Aldo Moro 2, 00185, Roma, Italy*
(Received 20 June 2005; published 31 August 2006)

The mixing properties of networks are usually inferred by comparing the degree of a node with the average degree of its neighbors. This kind of analysis often leads to incorrect conclusions: Assortative patterns may appear reversed by a mechanism known as Simpson's paradox. We prove this fact by analytical calculations and simulations on three classes of growing networks based on preferential attachment and fitness, where the disassortative behavior observed is a spurious effect. Our results give a crucial contribution to the debate about the origin of disassortative mixing, since networks previously classified as disassortative reveal instead assortative behavior to a careful analysis.

DOI: 10.1103/PhysRevE.74.026122 PACS number(s): 89.75.Hc, 89.75.Da, 89.75.Fb

Complex networks arise in a wide range of interacting structures, including social, technological, and biological systems [1]. Although these networks share some generic statistical features, such as the small-world property and, in many cases, the scale invariance of the degree distribution, they also display differences and peculiarities when their structure is examined in detail.

The mixing properties of a network refer to the attitude of nodes to connect to similar or unlike peers [2]. Similarity of nodes is established by comparing some node-dependent scalar quantity describing to a given quality. Networks where properties of neighboring nodes are positively correlated are called *assortative*, while those showing negative correlations are called *disassortative*. While assortativity, observed, for example, in social networks, finds an intuitive explanation in the fact that people (nodes) tend to build relations (links) with alike people, disassortativity is more puzzling and there is an open debate about its origins. A scalar quantity naturally associated to each node in a network is its degree, measuring the number of neighboring nodes. The mixing by degree (MBD) is often measured by looking at how the average degree K_{nn} of the nearest neighbors of a node depends on the degree K of the node itself. The mixing is assumed to be assortative when K_{nn} grows with K and disassortative when it decreases [3]. The relevance of MBD lies in that, beyond discriminating among different network morphologies [4,5], it reflects important structural properties: Assortative networks are found to be more resilient against the removal of vertexes than disassortative ones [6]. This implies, for example, that, when trying to block infection or opinion spreading within a social network [7] or to protect a computer network against cyber attacks [8], different strategies are needed depending on the MBD properties of the underlying network. It has recently been observed that the sign of degree correlations also affects other properties of complex networks such as synchronization [9].

Recent studies show that social networks exhibit assortative MBD, whereas technological and biological ones display disassortative MBD [10]. The Word Wide Web (WWW), a paradigmatic example of worldwide collaborative effort among millions of users and publishers, represents an anomaly: One would expect it to show assortative mixing, similarly to other social and collaborative networks, while it shows evidences of anticorrelations [2], and disassortative MBD [11].

This work focuses on the mixing properties of growing networks. Our aim is to demonstrate that even in the presence of genuine positive correlations between the degrees, spurious negative correlations may be observed. As a consequence, networks previously classified as disassortative reveal instead assortative behavior to a careful analysis. To characterize mixing patterns, one compares the degree of a node with the average degree of its neighbors. In growing networks a direction is naturally associated with the links; accordingly, each node has two kinds of neighbors: those to whom it links (downstream) and those linking to it (upstream). We show that distinguishing between the two kinds of neighbors when performing the averages is crucial: positive correlations between the degree of a node and the average degree of both upstream and downstream neighbors, considered separately, may be reversed when the degrees are averaged together ignoring the different nature of neighboring sites. The fact that pooling together data of different nature can reverse the results of a statistical analysis is well known in statistics and often encountered in social sciences, medical statistics, and finance. As counterintuitive as it may appear, this property contains no logical contradiction, although it is known in the literature as *Simpson's paradox* [12].

We show our result on two classes of complex growing networks: the linear preferential attachment (LPA) model [13] and the Bianconi–Barabási (BB) fitness model [14]. Both include as a special case the Barabási-Albert (BA) model [15]. In these networks links have a natural direction—from newly added nodes to existing ones. Thus, in the following we distinguish between upstream and downstream neighbors, respectively, along incoming and outgoing links.

To clarify our argument, we consider in detail the BA model (where calculations are simpler) before moving to the LPA and BB models. In the BA model, at each time step a node is added and attached to the network by m undirected links according to preferential attachment. A node i (introduced at time i) points to an existing node j with probability $p_j(i)$ proportional to its degree $K_j(i)$ at time i [15]. Although links are undirected in the original formulation, a direction

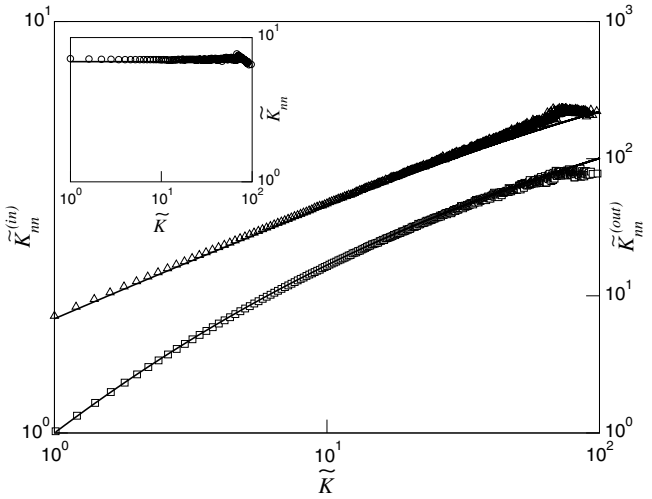

FIG. 1. $\widetilde{K}_{nn}^{(in)}$ (squares) and $\widetilde{K}_{nn}^{(out)}$(triangles) as functions of $\widetilde{K}$ from simulations of the BA model, with $m=100$, $t=10^4$, and averaged over 10^4 realizations. $\widetilde{K}_{nn}$ is shown in the inset. Solid lines represent the analytic calculations.

may be assigned to the links in a natural manner, from the newly added node to the one to preexisting ones. Since m sets a natural scale for the system, we will express all quantities in units of m and denote them by the superscript $\sim$. On average, the degree of node i grows in time as $\widetilde{K}_i(t) \simeq \sqrt{t/i}$ for $1 \ll i \ll t$ [15]. The average degrees of neighbors of i, in m units, read $\widetilde{K}_{nn,i}^{(in)}(t)=\Sigma_{j=i+1}^{t}\widetilde{K}_j(t)\widetilde{p}_i(j)/[\widetilde{K}_i(t)-1]$ and $\widetilde{K}_{nn,i}^{(out)}(t)=\Sigma_{j=1}^{i-1}\widetilde{K}_j(t)\widetilde{p}_j(i)$, where $\widetilde{K}_{nn,i}^{(in)}(t)$ and $\widetilde{K}_{nn,i}^{(out)}(t)$ refer to the degree of upstream and downstream neighbors, respectively. By approximating the sum by an integral and the degree by its average, one gets $\widetilde{K}_{nn,i}^{(in)}(t) \simeq \ln\sqrt{t/i}/(1-\sqrt{i/t})$ and $\widetilde{K}_{nn,i}^{(out)}(t) \simeq \sqrt{t/i}\,\ln(A\sqrt{i})$, where A is a constant of order 1 whose exact value depends on the initial condition [16]. At a given time t, we can express the above quantities in terms of $\widetilde{K}$ and drop the i dependence to get $\widetilde{K}_{nn}^{(in)} \simeq \widetilde{K}\ln\widetilde{K}/(\widetilde{K}-1)$ and $\widetilde{K}_{nn}^{(out)} \simeq \widetilde{K}\ln(\bar{K}/\widetilde{K})$, where $\bar{K}=A\sqrt{t}$ is of order of (and greater than) the maximum $\widetilde{K}$ observable at time t, $\widetilde{K}_{max} \simeq \sqrt{t}$. Thus $\widetilde{K}_{nn}^{(in)}$ is a monotonically (slowly) increasing function of $\widetilde{K}$, independent of t, and $\widetilde{K}_{nn}^{(out)}$ contains a t dependence through $\bar{K}$ and for any t is an increasing function of $\widetilde{K}$. We conclude that the degree of a node is positively correlated with the average degree of both upstream and downstream neighbors. However, computing the average degree of the neighbors altogether, correlations seem to vanish since one gets $\widetilde{K}_{nn}(t) \simeq \ln(A\sqrt{t})$, independent from $\widetilde{K}$ [17]. These results are confirmed by numerical simulation of the BA model and shown in Fig. 1, where histograms of $\widetilde{K}_{nn}^{(in)}$, $\widetilde{K}_{nn}^{(out)}$, and $\widetilde{K}_{nn}$ are plotted as functions of $\widetilde{K}$ for $t=10^4$ and $m=100$, averaged over 10^4 realizations.

Let us now focus on the LPA model [13], a generalization of the BA model: according to the same dynamics, at the ith time step m directed links are drawn from i to j with probability $p_j(i) \propto k_j(i)+\alpha$, where $k_j(i)$ is the in-degree of site j at time i. For $\alpha=m$, the BA model is recovered. When dealing with the LPA model, it is convenient to measure quantities in units α. In the continuous time limit, the time dependence of the in-degree is $\widetilde{k}_i(t)=(t/i)^{\beta}-1$ with $\beta=(1+\alpha/m)^{-1}$ [13]. The calculation of the average in-degree of upstream and downstream neighbors can be performed in analogy to the BA model (now $\widetilde{k}_{nn}^{(in)}$ and $\widetilde{k}_{nn}^{(out)}$count incoming links only). The average degree of upstream neighbors reads $\widetilde{k}_{nn}^{(in)} \simeq (\widetilde{k}+1)\ln(\widetilde{k}+1)/\widetilde{k}-1$, independent from the ratio α/m and thus coinciding with the result for the BA model, and is monotonically increasing. The average degree of downstream neighbors is given by

$$\widetilde{k}_{nn}^{(out)} \simeq \frac{1-\beta}{2\beta-1}(\widetilde{k}+1)\left[\left(\frac{\bar{k}+1}{\widetilde{k}+1}\right)^{(2\beta-1)/\beta}-1\right]-1,$$

where $\bar{k}(t)=A(\beta)^{\beta/(2\beta-1)}t^{\beta}-1$, with $A(\beta)=\frac{2\beta-1}{(1-\beta)^2}+2^{1-2\beta}$, and we have dropped the t dependence. Since $A(\beta) \geq 1$ for $\beta > 1/2$ and $0 \leq A(\beta) \leq 1$ for $\beta < 1/2$, we have $A(\beta)^{\beta/(2\beta-1)} \geq 1$ for any β. Thus $\bar{k}(t)+1 > \widetilde{k}_{max}+1 \simeq t^{\beta}$, where $\widetilde{k}_{max}$ is the maximum in-degree at time t measured in α units. One can check that in the limit $\alpha \to m$ $(\beta \to 1/2)$ this expression for $\widetilde{k}_{nn}^{(out)}$ coincides with that found for the BA model in the same continuous limit approximation. It is clear that $\widetilde{k}_{nn}^{(out)}$ is an increasing function of $\widetilde{k}$, both for $\beta > 1/2$ [where $\widetilde{k}_{nn}^{(out)}+1$ grows as $(\widetilde{k}+1)^{(1-\beta)/\beta}$] and for $\beta < 1/2$ [where $\widetilde{k}_{nn}^{(out)}+1$ grows as $(\widetilde{k}+1)$]. Thus, also for the LPA model, the in-degree of a node is positively correlated both with the average in-degree of incoming nearest neighbors and with the average in-degree of outgoing nearest neighbors. Since the in-degree differs from the degree by a constant, the same statement holds for the degree. Disregarding the different nature of the neighbors and averaging the degree over all nearest neighbors one gets

$$\widetilde{k}_{nn} \simeq \frac{\widetilde{k}+1}{\widetilde{k}+\beta/(1-\beta)}\left\{\ln(\widetilde{k}+1)+\frac{\beta}{2\beta-1}\times\left[\left(\frac{\bar{k}+1}{\widetilde{k}+1}\right)^{(2\beta-1)/\beta}-1\right]\right\}-1.$$

Two different regimes appear, for $\alpha/m < 1$ $(\beta > 1/2)$ and $\alpha/m > 1$ $(\beta < 1/2)$, separated by $\alpha=m$ where the LPA model coincides with the BA model. The average in-degree of nearest neighbors increases as a function of $\widetilde{k}$ for $\beta < 1/2$, while it decreases for $\beta > 1/2$ [18]. Thus, for the LPA model the average degree over all nearest neighbors increases or decreases for different values of the parameter β, even though the degree of a node is positively correlated with the average degree of both upstream and downstream nearest neighbors for any value of β [19]. In Figs. 2 and 3 we show the results of our calculation, compared with simulation of the LPA model for $m=100$, $t=10^4$, and $\alpha=5$ for the $\beta > 1/2$ regime (Fig. 2), and $m=100$, $t=10^4$, and $\alpha=500$ for the $\beta < 1/2$ regime (Fig. 3).

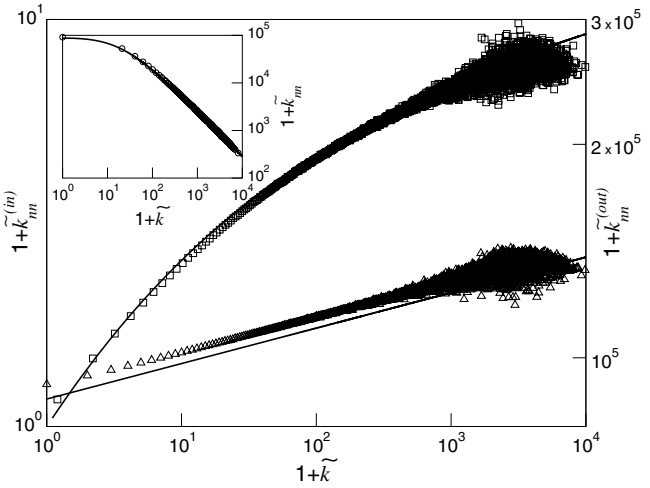

FIG. 2. $\tilde{k}_{nn}^{(in)}$ (squares) and $\tilde{k}_{nn}^{(out)}$ (triangles) as functions of $\tilde{k}$ for $t=10^4$ from simulations of the LPA model with $m=100$ and $\alpha=5$ ($\beta>1/2$) and averaged over 10^4 realizations. $\tilde{k}_{nn}$ is shown in the inset. Solid lines represent the analytic calculations.

Finally we consider the BB model [14], a paradigm for disassortatively mixed networks [3], originally proposed as a realistic model for the WWW. Here, the preferential attachment mechanism is modified to embody the intrinsic heterogeneity of nodes by assigning to each node j a quenched random variable, or *fitness*, η_j. The network is grown by adding a node at each time step and connecting it to m existing nodes chosen with probability proportional to both their degree and fitness $p_j(i+1)\propto\eta_j[k_j(i)+m]$. Now $k_j(i)$ depends on the realization of the network, and on the quenched variables $\{\eta_l\}_{l=1}^{i}$. However, for a given quenched disorder the degree is approximated by $k_j(t)\simeq m[(t/j)^{\eta_j/c}-1]$, where c depends on the probability distribution of the fitness and equals $1.255\ldots$ for a uniform distribution in $[0,1]$ (see the original paper by Bianconi and Barabási [14]). Thus $k_j(t)$ essentially

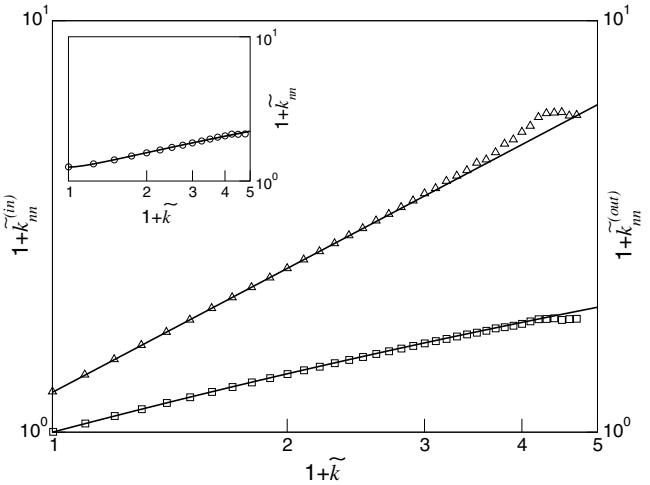

FIG. 3. $\tilde{k}_{nn}^{(in)}$ (squares) and $\tilde{K}_{nn}^{(out)}$ (triangles) as functions of $\tilde{k}$ for $t=10^4$ from simulations of the LPA model with $m=100$ and $\alpha=500$ ($\beta<1/2$) and averaged over 10^4 realizations. $\tilde{k}_{nn}$ is shown in the inset. Solid lines represent the analytic calculations.

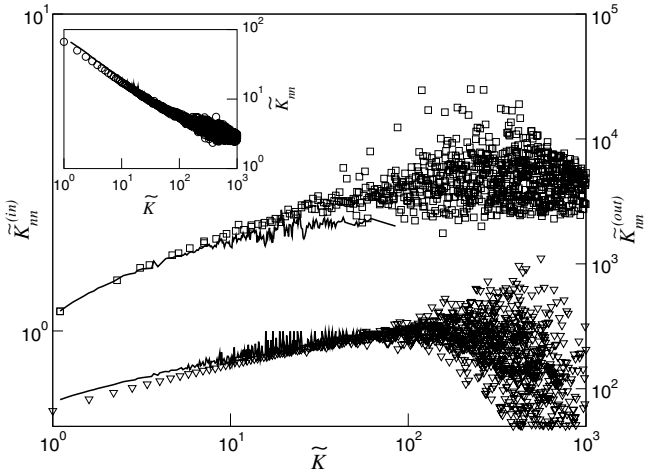

FIG. 4. $\tilde{K}_{nn}^{(in)}=\tilde{k}_{nn}^{(in)}+1$ (squares) and $\tilde{K}_{nn}^{(out)}=\tilde{k}_{nn}^{(out)}+1$ (triangles) as functions of $\tilde{K}=k+1$ for $t=10^4$ from simulations of the BB model with $m=10$ and averaged over 10^4 realizations. $\tilde{K}_{nn}=\tilde{k}_{nn}+1$ is shown in the inset. Solid lines represent the corresponding analytic expressions.

depends only on η_j. This approximation is found to be very accurate numerically, and we will use it in what follows. Also, we approximate $p_j(t)$ by replacing the normalization factor $\Sigma_{l=1}^{i}\eta_l[k_l(i)+m]$ with its average value mci [15]. In the same notations as above, we will measure quantities in units of m. The average degree of upstream neighbors is given by $\tilde{k}_{nn}^{in}(i,t,\eta_i)=\frac{\eta_i}{c\tilde{k}_i(t)}\Sigma_{j=i+1}^{t}\frac{\tilde{k}_i(j)+1}{j}\langle\tilde{k}_j(t)\rangle$; similarly, the average degree of upstream neighbors is $\tilde{k}_{nn}^{out}(i,t)=\Sigma_{j=1}^{i-1}\frac{\langle\tilde{k}_j(i)(\tilde{k}_j(t)+1)\rangle}{ci}$, where angular brackets represent the average over η_j for $j\neq i$, which yields

$$\tilde{k}_{nn,i}^{(in)}(t,\eta_i)=\frac{\eta_i}{ic\tilde{k}_i(t)}\sum_{j=i+1}^{t}(j/i)^{\eta_i/c-1}\frac{(t/j)^{1/c}-1}{\ln(t/j)^{1/c}},$$

$$\tilde{k}_{nn,i}^{(out)}(t)=\frac{1}{i}\sum_{j=1}^{i}[h((i/j)^2(t/i))-h(i/j)],$$

where $h(x)=\{x^{1/c}[\ln(x^{1/c})-1]+1\}/\ln^2(x)$. The $\tilde{k}$ dependence of $\tilde{k}_{nn}$ is then obtained by integrating out η_i:

$$\tilde{k}_{nn}^{(in)}(\tilde{k})=\frac{\int di\int d\eta\tilde{k}_{nn,i}^{in}(t,\eta)\delta(\tilde{k}_i(t)-\tilde{k})}{\int di\int d\eta\delta(\tilde{k}_i(t)-\tilde{k})},$$

which can be performed numerically. The results for a uniform distribution of fitness in $[0,1]$, confirmed by simulations, are shown in Fig. 4. Again, the degree of a node is positively correlated with the average degree of both upstream and downstream neighbors. However, as shown by Pastor-Satorras *et al.* [3], the nearest-neighbor average degree decreases as a function of the degree.

In conclusion, we have shown that in the analysis of mixing properties of growing networks genuine positive correlations can show up as negative through a reversal mechanism which is known as Simpson's paradox. This reversal mechanism affects the statistical analysis and typically takes place when two inhomogeneous populations are combined together (here, upstream and downstream neighbors). Several notable cases of occurrence have been recognized, for example, in the analysis of clinical trials [20]. In the growing network models examined in this paper, the counterintuitive effect of Simpson's paradox on the mixing properties can be understood as follows: the average neighbors' degree is larger on downstream than on upstream neighbors, although it grows with K in both cases. When computing the average neighbors' degree by mixing upstream and downstream nodes, one should carefully take into account how such mixture changes with K: as K increases, downstream neighbors contribute less and less to the total average since the out-degree of every node is a fixed number m. Accordingly, for small K the average neighbors' degree is determined mainly by downstream nodes (with large degrees) while the degree carried by upstream neighbors—typically smaller—dominates for large K. As a result, even though the average neighbors' degree grows with K when upstream and downstream neighbors are considered separately, the total average may be a flat function of K, as in the BA model case, or even a decreasing one (LPA model for $\beta<1$, BB model). Such behavior has often been used to infer disassortativity—for instance, in the WWW case [11,21]—since it suggests that nodes with many links are preferably linked to nodes with poor connectivity. Our work indicates that further analysis is needed to understand the mixing properties of real directed networks. In the case of the WWW, recent statistical analysis carried on of the in- (out-) degree–in (out-) degree correlations qualitatively confirm the presented scenario[22]. A similar behavior has also been observed in the analysis of software systems dependences [23].

We acknowledge useful discussions with M. A. Muñoz.

[1] For recent reviews, see R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002); M. E. J. Newman, SIAM Rev. **45**, 167 (2003); S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and Www* (Oxford University Press, Oxford, 2003); R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, England, 2004).

[2] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).

[3] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).

[4] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2005).

[5] G. Bianconi, G. Caldarelli, and A. Capocci, Phys. Rev. E **71**, 066116 (2005).

[6] A. Vázquez and Y. Moreno, Phys. Rev. E **67**, 015101(R) (2003).

[7] Ph. Blanchard, A. Krueger, T. Krueger, and P. Martin, e-print physics/0505031.

[8] Y. Hayashi and T. Miyazaki, e-print cond-mat/0503615.

[9] M. di Bernardo, F. Garofalo, and F. Sorrentino, e-print cond-mat/0506236.

[10] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).

[11] A. Capocci, G. Caldarelli, and P. De Los Rios, Phys. Rev. E **68**, 047101 (2003).

[12] E. H. Simpson, J. R. Stat. Soc. Ser. B. Methodol. **13**, 238 (1951); C. R. Blyth, J. Am. Stat. Assoc. **67**, 364 (1972).

[13] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).

[14] G. Bianconi and A.-L. Barabási, Europhys. Lett. **54**, 436 (2001).

[15] A. L. Barabási and R. Albert, Science **286**, 509 (1999).

[16] The average over realizations of the degree of a node i grows in time as $\tilde{K}_i(t)=\tilde{K}_i(i)f(t)/f(i)$, where $f(t)=\Gamma(t+1/2)/\Gamma(t)$. For large t this gives the asymptotic $\tilde{K}_i(t)\simeq\tilde{K}_i(i)c(i)\sqrt{t/i}$. Each node $i>1$ appears at time $t=i$ with a degree equal to m, so that $\tilde{K}_i(i)=1$. For $i=1$ it is convenient to choose $\tilde{K}_1(1)=2$, so that $\Sigma_{i=1}^{t}\tilde{K}_i(t)=2t$. The coefficient $c(i)$ is sensibly different 1 (value in the continuum time approximation) only for $i=1$, for which $c(1)=2/\sqrt{\pi}\simeq 1.128\ldots$, and we will approximate it with one for all other values, so that finally $\tilde{K}_i(t)\simeq\sqrt{t/i}$ for $i>1$, while $\tilde{K}_1(t)\simeq 4/\sqrt{\pi}\sqrt{t}$. With this initial conditions $A=e^{8/\pi}/\sqrt{2}$. Analogous calculations can be performed for the LPA model, where $f(t)$ is replaced by $\Gamma(t+\beta)/\Gamma(t)$. This gives $A=[2^{1-2\beta}+(2\beta-1)/(1-\beta)^2]^{\beta/(2\beta-1)}$.

[17] M. Boguñà and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).

[18] A. Barrat and R. Pastor-Satorras, Phys. Rev. E **71**, 036127 (2005).

[19] A behavior similar to the $\beta>1/2$ case was observed in simulations of a weighted directed model for the WWW: A. Barrat, M. Barthélemy, and A. Vespignani, Lect. Notes Comput. Sci. **3243**, 56 (2004).

[20] S. A. Julious and M. A. Mullee, Br. Med. J. **309**, 1480 (1994).

[21] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).

[22] M. A. Serrano, A. Maguitman, M. Boguna, S. Fortunato, and S. Vespignani (unpublished).

[23] C. R. Myers, Phys. Rev. E **68**, 046116 (2003).