Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations

Alexei Vázquez
Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556

(Received 23 November 2002; published 7 May 2003)

The linear preferential attachment hypothesis has been shown to be quite successful in explaining the existence of networks with power-law degree distributions. It is then quite important to determine if this mechanism is the consequence of a general principle based on local rules. In this work it is claimed that an effective linear preferential attachment is the natural outcome of growing network models based on local rules. It is also shown that the local models offer an explanation for other properties like the clustering hierarchy and degree correlations recently observed in complex networks. These conclusions are based on both analytical and numerical results for different local rules, including some models already proposed in the literature.

DOI: 10.1103/PhysRevE.67.056104 PACS number(s): 89.75.Hc, 89.75.Fb, 89.20.Hh, 89.65.−s

I. INTRODUCTION

In the last few years there has been a great interest in the study of networks, with particular emphasis on the following properties: small world effect [1,2], power-law degree distribution [3,4], and more recently degree correlations [5–7] and clustering hierarchy [6,8,9]. This explosion has been possible thanks to the increase of available network maps offering the graph representation for a wide variety of systems with sizes ranging from hundreds to billions of nodes. Examples include technological networks such as the physical Internet [5,6,10–17], the World Wide Web (WWW) [18–20], electronic mail [21,22], and electronic circuits [23], biological networks such as the protein-protein interaction network [24–28], metabolic paths [29,30], and food webs [31,32], and social networks represented by the citation graph [33–35], scientific collaboration webs [36–39], sexual relations [40], among others.

In particular, metrics like the degree (the number of edges incident to a vertex), the minimum path distance between pairs of vertices, and the clustering coefficient (the fraction of edges among the neighbors of a vertex) have attracted the attention of the physics community. Watts and Strogatz [1,2] have shown that, in general, real networks are characterized by a small average minimum path distance and a large clustering coefficient that together are named the *small world effect*. The name comes from the fact that we can reach every vertex in the graph by crossing a small number of edges. Moreover, Barabási and collaborators [41,42] have pointed out that many real networks are also characterized by power-law degree distributions, giving an appreciable probability to observe high-degree vertices. A more exhaustive analysis reveals that, in addition to power laws, truncated power laws and exponential distributions are also observed [43].

Barabási and Albert (BA) proposed a mechanism that explains the origin of power-law degree distributions [41]. This mechanism is based on two fundamental properties of a wide class of real networks, their growing nature, and the existence of a preferential attachment: new vertices added to the graph are attached preferentially to high-degree vertices. In particular, a linear preferential attachment, where the probability to get connected to a vertex is proportional to its degree, leads to power-law degree distributions. The preferential attachment mechanism can be generalized in different ways. A sublinear preferential attachment leads to bounded degree distributions while a superlinear one yields graphs with a single hub connected to almost any other vertex [44,45]. The power laws can also be truncated after the introduction of other ingredients such as aging [46], bounded capacity [43], or limited information [47]. Moreover, the introduction of quenched [48] and annealed [49,50] disorder leads to logarithmic corrections and multifractal scaling, respectively.

The BA model provides a general mechanism to obtain power-law degree distributions in growing networks. If one consider other measures like the clustering coefficient then one may conclude that this model is still insufficient to describe real graphs. However, we should not focus on the detailed properties of the model but on its philosophy. That is, if we assume that there is a growing tendency of the network and an effective linear preferential attachment then we obtain a scale-free degree distribution. Actually, this effective preferential attachment has been measured in different real graphs, including the Internet [5,51] and a variety of scientific collaboration webs [39,51,52], supporting the hypothesis of a linear attachment rate. With regard to the other topological properties, we can construct many models with different clustering coefficients, minimum path distances, and other metrics [53]. However, the origin of the ubiquity of the linear preferential attachment is not clear yet.

The topology of real networks is also characterized by degree correlations [5,7] and clustering hierarchy [6,9]. Moreover, these correlations influence the behavior of models defined on top of these graphs, as has recently been shown in Refs. [7,54–58]. Growing network models with global evolution rules, like the BA model, exhibit degree correlations. For instance, nontrivial degree correlations has been obtained in the linear preferential attachment model [45] and in a growing network model without any preferential attachment [59]. However, the degree correlations obtained in those global models are not sufficiently strong to account for the features observed in real graphs. New models giving a better representation of real graphs are starting to emerge [9,60,61]. In addition to the numerical simulations some analytical treatments have shown that power-law degree distributions and clustering hierarchy are obtained as an outcome of these models [9,62–66]. However, a general principle based on local rules is still missing.

In this work, different *local* mechanisms that lead to graphs with power-law degree distributions, degree correlations, and clustering hierarchy are studied. The term *local* means that we will investigate evolution rules that involve a vertex and its neighbors. As will be shown, the preferential attachment, the inverse proportionality between the average clustering coefficient and the vertex degree, and degree correlations are common features of growing graph models built by local rules. The general principles behind these features are also determined.

The paper is organized as follows. In the next section the motivation for this work is presented. It is shown that, in addition to power-law degree distributions, clustering hierarchy and degree correlations are common features of real networks. Then in the following sections three different models based on local rules are presented. In all cases both analytical and numerical evidence is provided. In particular, in Sec. III a walk model is proposed as a mechanism for searchable networks such as the WWW and the citation network. Then in Sec. IV a model for social network evolution is analyzed, based on the existence of potential connections between the neighbors of a vertex. Finally, in Sec. V we study models with duplication or replication of their vertices. The common patterns observed on these models are summarized in the concluding Sec. VI.

II. CORRELATIONS AND HIERARCHY IN REAL GRAPHS

In this section we study correlations in some real graphs. In particular we consider five different networks here denoted by Router, AS, WWW, Gnutella, PIN, and Math. In all cases the graph is obtained by representing the ‘‘relevant’’ units of the system by vertices and their interactions or relations by edges. In some cases, multiple graph representations of the same system can be obtained. ‘‘Router’’ is the router level graph representation of the Internet, where each vertex represents a router and each edge represents a physical connection among them. AS is the *autonomous system* (AS) representation of the Internet, where each vertex represents an AS or service provider and each edge represents a peer relation among them. WWW is the graph representation of the WWW, where each vertex represents a web page and each directed edge a hyperlink from one page to another. Here we will consider the directed edges as undirected. Gnutella is the graph representation of the peer-to-peer network of the same name, where each vertex represents a user and each edge a peer relation among them. PIN is the graph representation of the protein interaction network, where each vertex represents a protein and each edge an interaction among them. Math is the graph representation of the mathematical coautorship network, where each vertex represents an author and each edge the existence of at least one common publication among them.

In general, real networks are correlated and correlations may have different origins. Let us consider the example of the Internet. Due to installation costs, the Internet has been designed with a hierarchical structure. This hierarchy can be schematically divided into international connections, national backbones, regional networks, and local area networks. Vertices providing access to international connections or national backbones are of course on the top level of this hierarchy, since they make possible the communication between regional and local area networks. Moreover, in this way, a small average minimum path distance can be achieved with a small average degree. This hierarchical structure will introduce some correlations in the network topology. For instance, it is expected that vertices with high degrees are connected to vertices with low degrees.

In contrast, in social networks well connected people tend to be connected with well connected people [7]. Let us take the example of the scientific coauthorship graph. A scientist writing a lot of papers has in general a larger probability of writing a paper with another scientist who also has a lot of papers than with one with a few papers. In fact, if F_i is the number of papers of scientist i and $F_i \ll N$, then the probability that two scientists write a paper together is roughly F_iF_j/N. Now, F_i is in general a monotonically increasing function of the scientist degree d_i (number of collaborators) and, therefore, scientists with a high degree will have a better chance of making a new article together, i.e., of being connected.

To investigate these correlations it has been proposed to analyze the clustering coefficient and the nearest-neighbor average connectivity as a function of the vertex degree [5,6]. The clustering coefficient is the average probability that two neighbors l and m of a vertex i are connected. In terms of the adjacency matrix ($J_{ij}=1$ if vertices i and j are connected and 0 otherwise), the clustering coefficient is defined as the conditional probability that if $J_{il}J_{im}=1$ then $J_{lm}=1$. Thus, it measures in some way the existence of three-point correlations in the adjacency matrix. The clustering coefficient c_i is then defined as the ratio between the number of edges e_i among the d_i neighbors of a given vertex i and its maximum possible value, $d_i(d_i-1)/2$, i.e.,

$$c_i=\frac{2e_i}{d_i(d_i-1)}. \qquad (1)$$

The average clustering coefficient $\langle c\rangle$ is the average of c_i over all vertices in the graph. It provides a measure of how well the neighbors of a vertex are locally interconnected. In Refs. [1,2] it was shown that the clustering coefficient of many graphs representing real systems is orders of magnitude larger than the one expected for a random graph and, therefore, they are far from being random. Further information can be extracted if one computes it as a function of the vertex degree [6].

In Fig. 1 we plot $\langle c\rangle_d$ vs d for different real networks. According to this measure, two different classes emerge. In the first class (Math and Router data), $\langle c\rangle_d$ does not exhibit a strong dependency on d, except for finite size effects at the largest degrees. This behavior is typical of random graphs, where the probability that two neighbors of a vertex are connected by an edge is a constant, and equal to the probability that any two vertices selected at random are connected. On

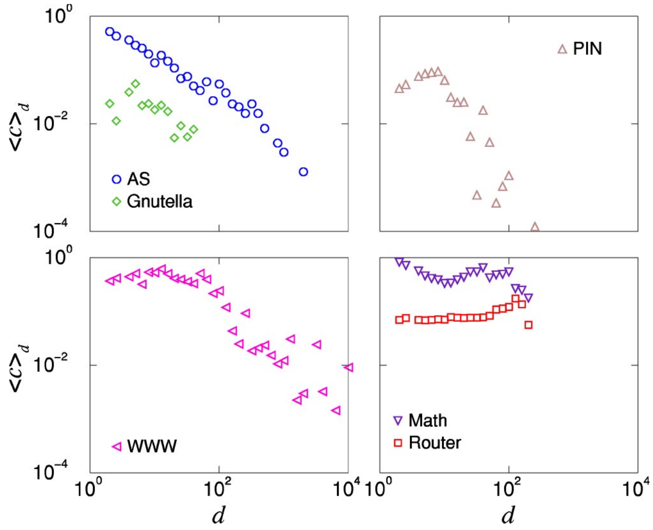

FIG. 1. Clustering coefficient as a function of the vertex degree for some real graphs. AS and Router are the autonomous system [10] and router [12] level graph representations of the Internet, respectively. WWW is a subgraph of the WWW network, a data set collected by the Notre Dame Group of Complex Networks [93]. Gnutella is the Gnutella peer-to-peer network, provided by Clip2 Distributed Search Solutions. PIN is the protein-protein interaction graph of *Saccharomices cerevisiae* as obtained from two hybrid experiments [26]. Math is the coauthorship graph obtained from all relevant journals in the field of mathematics and published in the period 1991–1998 [39].

the contrary, there is another class where $\langle c\rangle_d$ follows an evident decay with increasing vertex degree d. Thus, in this case, low-degree vertices form local subgraphs that are well connected. At the same time they are connected to other parts of the graph by high-degree vertices, having a few edges between the subgraphs they connect but giving a small average minimum path distance. This picture makes evident the existence of some hierarchy [5,6] or modularity [9].

These observations for the clustering coefficient are complemented by another metric related to the correlations between vertex degrees. These correlations are quantified by the probability $p(d'|d)$ that a vertex with degree d has an edge to a vertex with degree d'. With the available data a plot of this magnitude is very noisy and difficult to interpret. Thus in [5] it was suggested to measure the average degree among the nearest neighbors of a vertex, which is given by

$$\langle d_{nn}\rangle_d = \sum_{d'} d' p(d'|d), \tag{2}$$

and to plot it as a function of the vertex degree d. If there are not degree-degree correlations then the probability that an edge points to a vertex of degree d' is independent of d and proportional to $d' p_{d'}$, resulting, after normalization, in $p(d'|d) = d' p_{d'}/\langle d\rangle$. Therefore, the plot $\langle d_{nn}\rangle_d$ vs. d will be flat and equal to

$$\langle d_{nn}\rangle_{\text{unco}} = \frac{\langle d^2\rangle}{\langle d\rangle}. \tag{3}$$

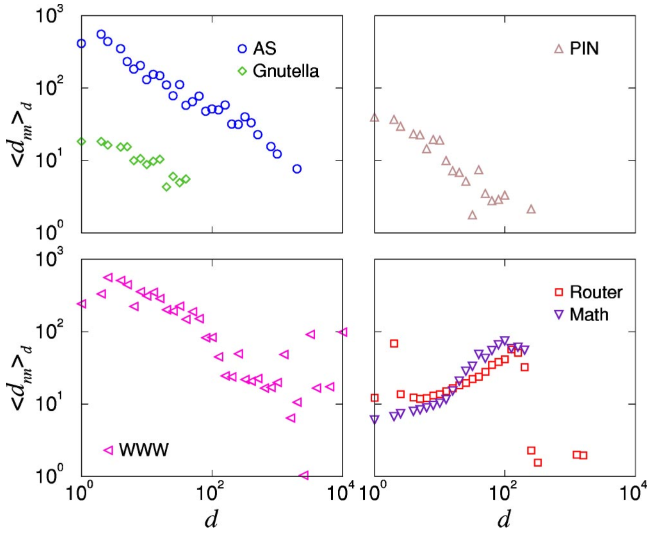

FIG. 2. Average nearest-neighbor degree as a function of the vertex degree for the real graphs introduced in Fig. 1.

In Fig. 2 we plot $\langle d_{nn}\rangle$ vs d for several real networks. In this case also we found the emergence of two different classes of graphs. In one of them the average nearest-neighbor degree exhibits a power-law decay with increasing vertex degree. This is strong evidence for the existence of disassortative (or negative) correlations, where large degree vertices tend to be connected with low-degree ones and vice versa. On the other hand, for some of the graphs (Math and Router data) an increasing tendency is observed, denoting the presence of assortative (or positive) correlations, where the edges connect vertices with similar degrees. The same conclusions are obtained using the Pearson coefficient of the degrees at either ends of an edge [7,67]. Notice that the subdivision attending either the clustering coefficient or the average nearest-neighbor degree coincides.

These observations cover a wide range of networks and are complemented by Refs. [5–7,9,67]. However, their origin is not yet clear. After some years of intensive research on complex networks there is no explanation for the ubiquity of the linear preferential attachment. Different models have been proposed but a mechanism is still missing. The lack of a general principle is extended to these new metrics associated with correlations. In the following sections three different models that exhibit these properties are studied, emphasizing the mechanism behind them. Based on their analysis some general conclusions will be achieved.

III. RANDOM WALK ON A NET

In this section we study the evolution of a graph where we know about new vertices by simply exploring the graph, with applications to searchable networks such as the citation and WWW graphs. We focus on different local mechanisms, where the term “local” means that we will investigate evolution rules that involve a vertex and its neighbors. A global approach based on effective attachment rates can be found in [68].

There are different ways to obtain information about the documents (articles, web pages) in these graphs, like looking at directories (citation index, web crawler), commercial spots, shown by a friend, or following the references (citations, hyperlinks) that are contained in the documents that we already know. In the case of the citation graph, we often find new articles from the citation list of an article that we already know and, later on, we can repeat the process with these new articles. Moreover, it is known that with a high probability people know about new web pages by surfing on the WWW.

Two of the major contributions to how people find out about new web pages are following the hyperlinks of other web pages and using search engines [69]. The first source can be characterized by modeling the WWW "surfers" as random walkers on the WWW graph. Let us assume that the walk starts from a page selected at random and, on each page, with probability q_e it decides to follow one link on that page or to jump to another random page with probability $1-q_e$. Then, the probability v_i that a page i will be visited is given by

$$v_i=\frac{1-q_e}{N}+q_e\sum_j J_{ij}\frac{v_j}{d_j^{ou}}, \tag{4}$$

where J_{ij} is the adjacency matrix and d_j^{ou} denotes the vertex out degree. It is quite interesting to notice that this probability of being visited by a random surfer is often used by search engines as a page rank criterion [70], as is the case with the popular Google [71]. Hence, the two main sources through which new pages are visited are characterized by Eq. (4) and, therefore, the main properties of the in-degree distribution of the WWW graph should be computed by starting on it. However, to my knowledge and except from the recursive search model proposed by the author in Ref. [72], no study has been performed in this direction.

In a mean-field approximation one can replace the sum in Eq. (4) by Θd_i^{in}, resulting in

$$v_i=\frac{1-q_e}{N}+q_e\Theta d_i^{in}, \tag{5}$$

where Θ is the average probability that a vertex pointing to vertex i is visited and d_i^{in} is the vertex in degree. To compute Θ we should take into account that the probability that a vertex i has an in edge coming from a vertex with out degree d^{ou} is $d^{ou}p_{d^{ou}}/\langle d^{ou}\rangle$. This edge will be selected at random among the d^{ou} out edges and, therefore, with probability $1/d^{ou}$. Thus,

$$\Theta=\sum_{d^{ou}}\frac{d^{ou}p_{d^{ou}}}{\langle d^{ou}\rangle}\frac{1}{d^{ou}}v_{d^{ou}}=\frac{\langle v\rangle}{\langle d^{ou}\rangle}. \tag{6}$$

In general when we visit new pages we do not create a hyperlink to it. In a first approximation this can be modeled by introducing the probability q_v that a visited vertex (page) increases its in degree by 1 (a hyperlink is created to it). Then, when a walk is performed $\langle v\rangle N$ vertices are visited and, therefore, $q_v\langle v\rangle N$ edges are added on average, resulting in

$$\frac{\partial N}{\partial t}=\nu_a,$$

$$\frac{\partial E}{\partial t}=\nu_s q_v\langle v\rangle N, \tag{7}$$

where E is the number of edges, and ν_s and ν_a are the number of surfers and the number of newly added pages per unit time, respectively. The integration of these equations yields

$$\langle d^{ou}\rangle=\langle d^{in}\rangle=q_v\langle v\rangle N\frac{\nu_s}{\nu_a}. \tag{8}$$

Thus, from Eqs. (6) and (8) we finally obtain

$$\Theta=\frac{\nu_a}{q_v\nu_s N}. \tag{9}$$

The probability that the in degree of a vertex of in degree $d^{(in)}$ increases by 1 when a surfer walks on the graph is given by $A(d^{(in)})=q_v v(d^{(in)})$ and, therefore, from Eqs. (5) and (9) it follows that

$$A(d^{(in)})=\frac{1}{N}\left[q_v(1-q_e)+q_e\frac{\nu_a}{\nu_s}d^{(in)}\right]. \tag{10}$$

Notice that the walk on the graph leads to an effective linear preferential attachment. The degree distribution corresponding to this attachment rate can easily be obtained using the rate equation approach [44,45]. Indeed, the number of vertices $n_{d^{in}}(t)$ with in degree d^{in} satisfies the rate equation

$$\frac{\partial n_{d^{in}}}{\partial t}=\nu_s A_{d^{in}-1}n_{d^{in}-1}-\nu_s A_{d^{in}}n_{d^{in}}+\nu_a\delta_{d^{in}0}. \tag{11}$$

Now we should take into account that the number of vertices on the WWW graph grows exponentially and, in this case, $\nu_a\propto N$. Moreover, assuming that each surfer has its own (or group of) web page (pages) the number of surfers is expected to be proportional to the number of web pages, i.e., $\nu_s\propto N$. Thus,

$$\frac{\nu_s}{\nu_a}=\alpha, \tag{12}$$

where α is a constant. It is worth noticing that Eq. (12) is always satisfied for networks with a constant growth rate, as may be the case of the citation graph. If this condition is satisfied then the in-degree distribution reaches a stationary state and we can write $n_{d^{in}}(t)=Np_{d^{in}}$, where $p_{d^{in}}$ is the stationary probability that a vertex has in degree d^{in}. Substituting this expression in Eq. (11), we obtain

$$p_{d^{in}}=\frac{1}{1+a}\,\frac{\Gamma[a(\gamma-1)+d^{in}]}{\Gamma[a(\gamma-1)]}\,\frac{\Gamma[(1+a)(\gamma-1)+1]}{\Gamma[(1+a)(\gamma-1)+d^{in}+1]},\tag{13}$$

where

$$\gamma=1+\frac{1}{q_e},\quad a=\alpha q_v(1-q_e)\tag{14}$$

with the asymptotic behavior for large in degree

$$p_{d^{in}}\sim(d^{in})^{-\gamma}.\tag{15}$$

Hence, the random walk model on a directed graph leads to a power-law in-degree distribution, with an exponent $\gamma \geqslant 2$. Notice that the power-law exponent does not depend on q_v and, therefore, we expect that generalizations of the rule of creating an edge to a visited vertex will not change this exponent. For instance, one can divide the vertices into classes in such a way that the edges can be created only among vertices of the same class, and the resulting power-law exponent should be the same. Moreover, the power-law exponent does not depend on α.

We can go beyond the in-degree distribution and compute the clustering coefficient as a function of the total degree $d = d^{in}+d^{ou}$ of a vertex. For this purpose we consider the graph as undirected and compute the number e_i of edges among the neighbors of a vertex i. Since the only dynamics in this model is given by the random walk, the result is

$$\frac{\partial e_i}{\partial t}=q_v(q_e\Theta d_i^{in}+q_e v_i).\tag{16}$$

The first term on the right-hand side is the probability that a vertex with an out edge to i is visited and the second the probability that vertex i is visited and the walk follows one of its out edges to visit an out-neighbor vertex. In all cases the visited vertex is selected with probability q_v. Using Eqs. (5), (9), and (10) and taking into account that $\partial_t d_i^{in} = A(d_i^{in})$, we can rewrite Eq. (16) as

$$\frac{\partial e_i}{\partial t}\approx(1+q_e)\frac{\partial d_i^{in}}{\partial t},\tag{17}$$

where we have neglected the first term in the right-hand side of Eq. (10). Integrating this equation with the boundary condition $e(d^{in}=0)=0$ we obtain the clustering coefficient

$$\langle c\rangle_d=\frac{2e(d)}{d(d-1)}=\frac{2(1+q_e)}{d}+\frac{2(1+q_e)(1-d^{ou})}{d(d-1)}.\tag{18}$$

For large d the clustering coefficient scales as

$$\langle c\rangle_d\approx\frac{2(1+q_e)}{d}.\tag{19}$$

Thus, we obtain an inverse proportionality between the clustering coefficient and the vertex degree.

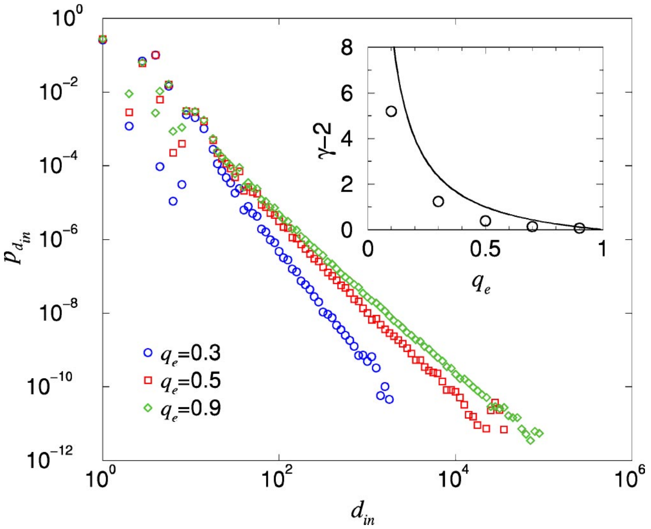

FIG. 3. In-degree distribution of the random walk model for different values of the probability of continuing the walk q_e and for graph size $N=10^6$. In all cases we take the average over 100 realizations. The inset shows the exponent γ obtained from the fit to the power law $p_{d^{in}}\sim(d^{in})^{-\gamma}$ (circles) together with the analytical prediction (continuous line).

A. Random walk model

We now study a particular random walk model by means of numerical simulations and compare its properties with the analytical results obtained above. We have made some simplifications in order to reduce the number of parameters and investigate the influence of the most important parameter q_e. The model is defined as follows. *Initial condition*: we start with one vertex and an empty set of edges. Then we iteratively perform the following rules.

Adding. A new vertex is created with an edge pointing to one of the existing vertices, which is selected at random.

Walking. If an edge is created to a vertex in the network then with probability q_e an edge is also created to one of its nearest neighbors. When no edge is created, go to the *adding* rule.

The first simplification is that there is only one "surfer" in the network, i.e., $\nu_s=1$. Second, each time the "surfer" decides not to follow one of the edges of the visited vertex it stops, and a new vertex starts a search from a vertex selected at random. In other words, the jump to a random vertex is coupled with the addition of new vertices resulting in $\nu_a = 1-q_e$. Finally, each time a vertex is visited an edge is created to it; thus $q_v=1$. Hence, the in-degree distribution is given by Eq. (13) with

$$\gamma=1+\frac{1}{q_e},\quad a=1.\tag{20}$$

We have made numerical simulations of this random walk model up to graph sizes $N=10^6$ taking an average over 100 realizations. In Fig. 3 we show a log-log plot of the in-degree distribution for different values of q_e. The power-law decay for large in degrees is evident. The exponent γ obtained from the fit to the numerical data is shown in the inset, together

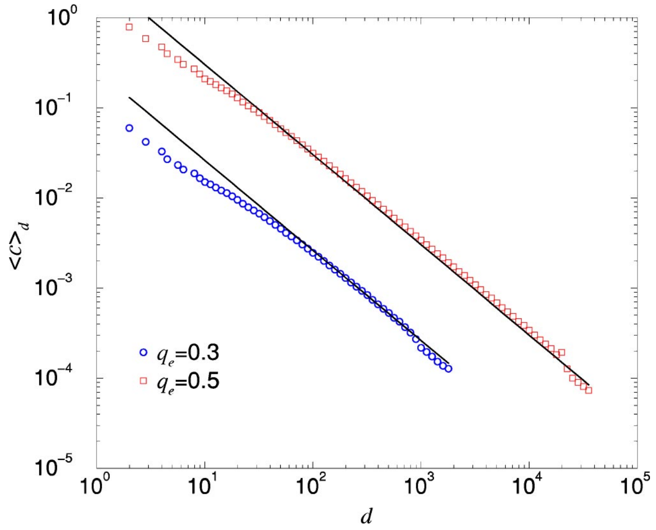

FIG. 4. Clustering coefficient as a function of vertex degree of the random walk model, for different values of the probability of continuing the walk q_e and for graph size $N=10^6$. In all cases we take the average over 100 realizations. The solid lines correspond to the power-law decay $C(d)=2(1+q_e)/d$.

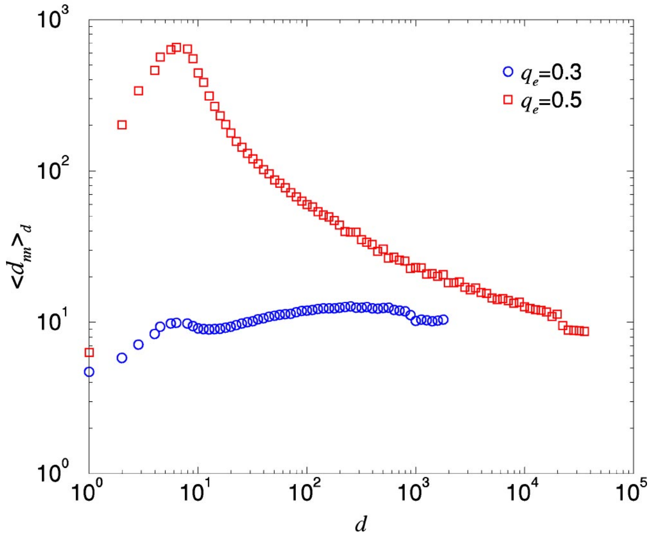

FIG. 5. Average neighbor degree as a function of vertex degree of the random walk model, for different values of the probability to continue the walk q_e and for graph size $N=10^6$. In all cases we take the average over 100 realizations.

with the predicted dependency in Eq. (20). The analytical values overestimate the power-law exponent but the qualitative picture is the same. For $q_e\to 0$ the power-law exponent is so large that the degree distribution cannot be distinguished from an exponential distribution. In contrast, for $q\to 1$ it approaches is minimum value $\gamma=2$. We attribute the quantitative disagreement to the mean-field approximation performed in the step from Eq. (4) to Eq. (5). On the other hand, the behavior of the average clustering coefficient with respect to the vertex degree is shown in Fig. 4. In this case the analytical asymptotic behavior in Eq. (19) is in very good agreement with the numerical data.

We were not able to obtain a prediction for the scaling of the average neighbor degree with the vertex degree. In this case our analysis relies on numerical simulations. In Fig. 5 we plot $\langle d_{nn}\rangle$ vs d for two values of q_e. For $q_e=0.3$ and for small values of q_e the average neighbor degree does not exhibit a strong dependency on d and, therefore, the graph appears uncorrelated. In contrast, for $q_e=0.5$ and in general for larger values of q_e it shows a peak around $d=10$ and then decays with increasing degree. This decay becomes even faster with increasing q_e. We have not found an explanation for this qualitative change of behavior yet. It is worth noticing that the experimental data for the WWW yield $\gamma\approx 2.1$, which can be obtained with our model using $q_e>0.5$. For this value of q_e the model yields negative correlations in agreement with the real data presented in Sec. II. However, we should take into account that the above analysis includes the fluctuation properties of the in degree, while the statistics of the out degree was not considered. The last one is irrelevant to determining the in-degree distribution but has to be taken into account to determine the clustering and degree correlation properties of the undirected representation of the directed graph. Hence, the results obtained here for $\langle c\rangle_d$ and $\langle d_{nn}\rangle_d$ are not conclusive.

B. Recursive search model

In the random walk model one follows only one edge of the visited vertices. However, one may consider an exhaustive search following all the edges recursively [72]. The main idea of a recursive search is thus to be connected to one vertex of the network, and any time we get in contact with a new vertex we follow all its edges, exploring in this way a larger part of the network. This can be modeled by modifying the walking rule as follows.

Walking. If an edge is created to a vertex in the network then with probability q_e an edge is also created to each of its nearest neighbors. When no edge is created go to the *adding* rule.

As for the previous model we have $\nu_s=1$, $\nu_a=1-q_e$ but $A(d^{in})$ is not given by Eq. (10). The form of $A(d^{in})$, and consequently the in-degree distribution, is determined below for two limiting cases.

q_e=0. In this case only the *adding* rule is performed; hence $A(d^{in})=1/N$ independent of d^{in}. The fact that $A(d^{in})$ scales as N^{-1} carries as a consequence that $n_{d^{in}}(N)=Np_{d^{in}}$ is the stationary solution of Eq. (11), where $p_{d^{in}}$ is the stationary probability of finding a vertex with in degree d^{in}. Substituting this expression in Eq. (11), one obtains

$$p_{d^{in}}=2^{-(d^{in}+1)}. \tag{21}$$

q_e=1. For this limiting case also the in-degree distribution can be computed exactly. Let us determine $A(d^{in})$ using the following fact. Any vertex i with in degree d_i^{in} has d_i^{in} vertices with an edge to it, which will be denoted by $x_j(j=1,2,\dots,d_i^{in})$. At the same time each of these x_j vertices may have other vertices with an edge to it. The following result holds: Any vertex with an edge to any of the vertices x_j also has an edge to i. The proof is straightforward. If when a vertex is added it creates an edge to any of the vertices x_j

then with probability $q_e=1$ it creates an edge to all the nearest neighbors of x_j, among which the vertex i is contained; end of proof. Hence, the probability that when a vertex is added it creates an edge to vertex i is just the probability $(1+d_i^{in})/N$ that the first edge is connected to i or to any of the d_i^{in} vertices with an edge to i, i.e., $A(d^{in})=(1+d^{in})/N$. As for $q_e=0$ $A(d^{in})$ scales as $1/N$ and, therefore, the stationary solution is of the form $n_{d^{in}}(N)=Np_{d^{in}}$. Then from Eq. (11) it follows that

$$p_{d^{in}}=\frac{1}{(d^{in}+1)(d^{in}+2)}. \tag{22}$$

Notice that in this case also, although it is not implicitly assumed, there is a preferential attachment leading to the power-law decay for large in degrees $p_{d^{in}}\sim(d^{in})^{-2}$.

The limiting cases $q_e=0$ and $q_e=1$ are described by in-degree distributions which are qualitatively different. For $q_e=0$ the distribution is exponential with a finite average in degree. In contrast, for $q_e=1$, the distribution follows a power-law decay $p_{d^{in}}\sim d^{in-\gamma}$ for large d^{in}, with $\gamma=2$. This power-law decay goes up to the largest possible degree $d^{in}\sim N^{1/(\gamma-1)}\sim N$ while $p_{d^{in}}=0$ for $d^{in}\geqslant N$. Hence, for $q_e=1$ and large N the average in degree scales as

$$\langle d^{in}\rangle(N)=\langle d^{ou}\rangle(N)=a+\ln N, \tag{23}$$

where a is independent of N and clearly $\langle d^{in}\rangle$ diverges in the thermodynamic (large network sizes) limit. In a mean-field approximation one can neglect the existence of loops in the network and, in such a case, the ‘‘walking’’ rule will take place on a tree. Each vertex on the tree will have on average $\langle d^{ou}\rangle(N)$ sons, which is just the average out degree after N vertices have been added. Moreover, if a vertex is visited then each of its sons will be visited with probability q_e. Hence, when the vertex $N+1$ is added, its average out degree $\langle d^{ou}\rangle(N+1)$ will be given by the average number of vertices visited during the walk, i.e.,

$$\langle d^{ou}\rangle(N+1)=1+q_e\langle d^{ou}\rangle(N)+[q_e\langle d^{ou}\rangle(N)]^2+\cdots$$

$$=\frac{1}{1-q_e\langle d^{ou}\rangle(N)}. \tag{24}$$

If there is a stationary state then $\langle d^{ou}\rangle(N+1)=\langle d^{ou}\rangle(N)=\langle d^{ou}\rangle$. In this case Eq. (24) yields two solutions. One of them diverges when $q_e\rightarrow 0$, which is not admissible since $\langle d^{ou}\rangle=1$ for $q_e=0$. The other solution reads

$$\langle d^{ou}\rangle=\langle d^{in}\rangle=\frac{1-\sqrt{1-4q_e}}{2q_e}. \tag{25}$$

This solution is valid for $q_e\leqslant q_c=1/4$ and, therefore, the average out degree does not converge to a stationary value when $q_e>q_c$. In this last region the average out degree increases logarithmically with N, as in the extreme case $q_e=1$ [see Eq. (23)]. Now, $\langle d^{in}\rangle=\langle d^{ou}\rangle$ and both approach a stationary state for any $\gamma>2$ and diverge otherwise. We then expect that the in-degree distribution has a power-law exponent $\gamma>2$ for $q_e<q_c$ and $\gamma\leqslant 2$ for $q_e>q_c$. Moreover, taking into account that the fastest divergence is obtained for $q_e=1$, where $\gamma=2$, we conclude that for $q_e>q_c$ the power-law exponent is constant and equal to $\gamma=2$.

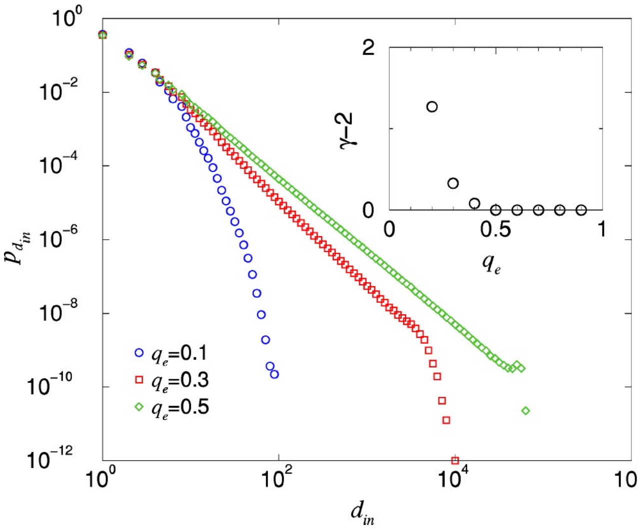

FIG. 6. Log-log plot of the in-degree distribution of the recursive search model for different values of q_e. The inset shows the exponent γ obtained from the power-law fit $p_{d^{in}}\sim(d_{in})^{-\gamma}$ to the numerical data.

To investigate the behavior for $0<q_e<1$ and the existence of a nontrivial threshold q_c as predicted by the mean-field approach, we have made numerical simulations of the recursive search model for different values of q_e up to graph sizes $N=10^5$. For each value of q_e the in-degree distribution was averaged over 100 runs of the algorithm. The resulting in-degree distribution is shown in Fig. 6. For $q_e=0.1$ the decay for large in degrees is very fast, and can be fitted by a power-law decay with a very large exponent, or equivalently by an exponential decay. On the contrary, for larger q_e the exponent becomes smaller and the power-law behavior becomes more evident. Finally, for $q_e\geqslant q_c=0.5\pm 0.1$, the exponent becomes independent of q_e and equals $\gamma=2$, in agreement with the mean-field prediction. However, the numerical threshold is twice the value obtained from Eq. (25).

In ordinary critical phenomena there is an absence of any typical length scale at the critical point, which is observed at a precise value of the order parameter. For the present model, however, the absence of a characteristic in degree is manifested not only at a precise value of q_e but in the whole interval $q_c\leqslant q_e\leqslant 1$. These features are very similar to those observed in some sandpile models [73,74], the paradigm of self-organized critical systems [75,76]. As in these models [77,78], there is a time scale separation between the addition of new vertices and their ‘‘walk’’ through the network. In the thermodynamic limit $(N\rightarrow\infty)$ the phase diagram of the model is divided into a subcritical $(0\leqslant q_e<q_c)$ and a critical region $(q_c\leqslant q_e\leqslant 1)$, where the power-law exponent does not depend on the control parameter. Hence, the results presented here suggest that for $q_c\leqslant q_e\leqslant 1$ the present model is in a self-organized critical state.

IV. CONNECTING NEAREST NEIGHBORS

In social graphs it is more probable that two vertices with a common neighbor get connected than two vertices chosen at random [52]. Clearly, this property leads to a large average clustering coefficient since it increases the number of connections between the neighbors of a vertex, as has already been observed in a model proposed by Davidsen, Ebel, and Bornholdt (DEB) [79]. The basic assumption of their model is that the evolution of social connections is mainly determined by the creation of new relations between pairs of individuals with a common friend. Moreover, a similar mechanism was considered by Holme and Kim [61] and by Jin *et al.* [38] to introduce an appreciable clustering coefficient in preferential attachment models.

The study of these models has been mainly performed by numerical simulations. A deeper analytical understanding can be obtained by introducing the concept of potential edge. We will say that a pair of vertices is connected by a *potential edge* if (1) they are not connected by an edge and (2) they have at least one common neighbor. Notice that while this concept has been implicitly considered in previous work its mathematical description will be introduced here.

The graph dynamics will be defined by the transition rates between the three possible states of a pair of vertices: disconnected (s), on connected by a potential edge (p) or by an edge (e). Let d_i^* be the number of potential edges incident to vertex i, the potential degree, to abbreviate. We can write the rate equations for the evolution of the number of vertices with degree d and potential degree d^*. Instead we will use the continuum approach [80,81]. In this case we neglect fluctuations and write mean-field equations for the evolution of d_i and d_i^*,

$$\frac{\partial d_i}{\partial N} = \nu_{s\to e}\hat{d}_i + \nu_{p\to e} d_i^* - (\nu_{e\to s} + \nu_{e\to p}) d_i\,,$$

$$\frac{\partial d_i^*}{\partial N} = \nu_{s\to p}\hat{d}_i + \nu_{e\to p} d_i - (\nu_{p\to s} + \nu_{p\to e}) d_i^*\,,$$

$$\hat{d}_i = N - d_i - d_i^*\,. \tag{26}$$

$\nu_{x\to y}$ is the transition rate from state x to state y per unit of N and $\hat{d}_i$ is the number of remaining neighbors, which are not connected by a potential edge or by an edge to vertex i.

The creation (deletion) of a potential edge incident to a vertex is associated with the creation (deletion) of an edge incident to one of its neighbors. For instance, if a new vertex i is connected to an existing vertex j then a potential edge is created between i and all neighbors of j. Hence

$$\nu_{s\to p} = \nu_{s\to e} d_i\,,$$

$$\nu_{p\to s} = \nu_{e\to s} d_i\,. \tag{27}$$

These equalities are at the core of the connecting nearest-neighbor model.

In the following we will neglect any process where an edge is deleted, i.e.,

$$\nu_{e\to s} = 0. \tag{28}$$

This assumption may seem too crude for some social networks where it is known that social relations can be lost, but it is realistic in many other cases. For instance, in the network of scientific collaborations two scientists are said to be connected if they have coauthored a paper. It is clear that this connection cannot be lost in time because the fact that they have written a paper together cannot be changed. In general, if the connection between two vertices is given by the occurrence of a certain event (coauthoring a paper, being in the cast of the same film, having a sexual relation) in the past history, then this connection cannot be lost and, therefore, our approximation holds.

Another crucial assumption is related to the fact that the transition from a potential edge to an edge has a higher probability of occurrence than the transition from being disconnected to an edge. In fact, the connection of two disconnected vertices without a common neighbor is a process that models the creation of a social relation between two social entities chosen at random. We thus assume

$$\nu_{s\to e} = \frac{\mu_0}{N^2}. \tag{29}$$

On the other hand, the creation of an edge between two vertices with a common neighbor, that is, with a potential edge between them, models the creation of a social relation between two ‘‘friends’’ of a social entity. In this case we assume

$$\nu_{p\to e} = \frac{\mu_1}{N}. \tag{30}$$

Under these approximations the system of equations (26) is reduced to

$$N\frac{\partial d_i}{\partial N} = \mu_0 + \mu_1 d_i^*\,,$$

$$N\frac{\partial d_i^*}{\partial N} = \mu_0 d_i - \mu_1 d_i^*\,. \tag{31}$$

Hence, the existence of a linear preferential attachment (the growth rates of d_i and d_i^* are linear in themselves) in this class of models becomes evident with the introduction of the concept of potential edges. Thus, a power-law degree distribution is expected. This system of differential equations is linear and, therefore, can be easily integrated, with the result that, for $N \gg N_i$,

$$d_i(N) = d_0\left(\frac{N}{N_i}\right)^{\beta}, \quad d_i^*(N) = d_0^*\left(\frac{N}{N_i}\right)^{\beta}, \tag{32}$$

where N_i is the size of the graph when vertex i was added to it and

$$\beta = \frac{\mu_1}{2}\left(-1 + \sqrt{1 + 4\frac{\mu_0}{\mu_1}}\right). \tag{33}$$

Now, if the vertices are added at a constant rate then $P(N_i = N) = 1/N$, yielding

$$P(d_i > d) = P\left[d_0\left(\frac{N}{N_i}\right)^{\beta} > d\right]$$
$$= \int_0^N \frac{dN_i}{N}\Theta\left[d_0\left(\frac{N}{N_i}\right)^{\beta} - d\right]. \tag{34}$$

Consequently,

$$p_d = \frac{\partial P(d_i > d)}{\partial d} \sim d^{-\gamma} \tag{35}$$

with

$$\gamma = 1 + \frac{1}{\beta}. \tag{36}$$

Notice that the main ingredient leading to this power-law behavior is given by Eq. (27). In contrast, if $\nu_{s\rightarrow p}$ were independent of the vertex degree an exponential decay would be obtained.

We can also compute the clustering coefficient as a function of the vertex degree. The main contribution to the evolution of e_i, the number of edges among the neighbors of vertex i, is given by the transition *potential edge* → *edge*. In fact, if the potential edge connecting a vertex i to another vertex j, with common neighbor k, becomes an edge then vertex i gains one neighbor (vertex j) and a new edge among its neighbors (that connecting j and k). Neglecting other contributions we have

$$\frac{\partial e_i}{\partial N} = \nu_{p\rightarrow e} d_i^* = \mu_1 \frac{d_i^*}{N}. \tag{37}$$

Integrating this equation using Eq. (32), the result is

$$\langle c \rangle_d = \frac{2e(d)}{d(d-1)} \approx \frac{2\mu_1}{d}. \tag{38}$$

Thus, once again we obtain the inverse proportionality between $\langle c \rangle_d$ and vertex degree d, in this case due to the conversion of potential edges between vertices with a common neighbor into edges.

A. Connecting nearest-neighbor model

To check these results we have made numerical simulations of a variant of the DEB model. Starting with a single vertex and an empty set of edges iteratively perform the following rules.

(1) With probability $1-u$ introduce a new vertex in the graph, create an edge from the new vertex to a vertex j selected at random (implying the creation of a potential edge between the new vertex and all the neighbors of j).

(2) With probability u convert one potential edge selected at random into an edge.

A schematic representation of these rules is shown in Fig. 7. Actually, in the DEB model the number of vertices is fixed and each time a new vertex is added one vertex is removed from the graph. We consider the growing variant because in this case it is easier to determine some properties analytically. For very large N we expect that both variants have the same qualitative behavior.

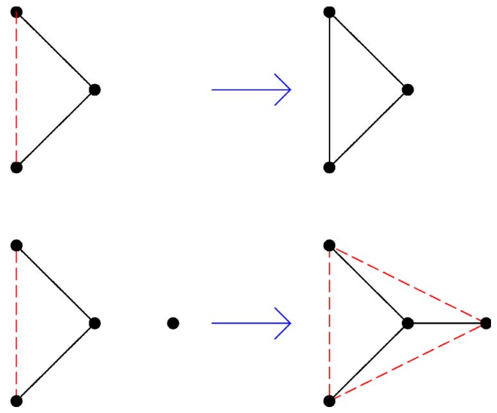

FIG. 7. Schematic representation of the two evolution rules of the connecting nearest-neighbor model. Top: with probability u a potential edge (dashed line) becomes an edge (continuous lines). Bottom: with probability $1-u$ a new vertex is added to the graph (disconnected vertex on the left); then it is connected with an edge to a vertex selected at random and by potential edges to its neighbors (right).

These evolution rules fit into the equations written above after setting

$$\mu_0 = 1, \quad \mu_1 = \frac{u}{1-u}. \tag{39}$$

Thus, from Eqs. (33) and (36) it follows that

$$\gamma(u) = 1 + \frac{2(1-u)}{u}\left(-1 + \sqrt{1 + 4\frac{1-u}{u}}\right)^{-1}, \tag{40}$$

with the limiting cases

$$\gamma(0) = \infty, \quad \gamma(1) = 2. \tag{41}$$

Thus, the power-law exponent γ takes its minimum value when $u \rightarrow 1$, corresponding to a low rate of addition of vertices, and it grows with decreasing u corresponding to higher rates of vertex addition. In Fig. 8 we plot the degree distribution as obtained from numerical simulations. For intermediate degrees it exhibits a power-law decay $p_d \sim d^{-\gamma}$. The value of γ obtained from the fit to the numerical data is shown in the inset, together with the analytical curve given by Eq. (40). The quantitative disagreement tells us that the mean-field Eq. (26) give us the right qualitative description but fluctuations should be considered to obtain a precise agreement with the numerical data.

In Fig. 9 we plot the clustering coefficient as a function of the vertex degree. It follows a power-law decay for large degrees but with an exponent smaller than 1. On the other hand, the average neighbor degree as a function of the vertex degree is shown in Fig. 10. It increases with increasing d, i.e., the graphs generated using this model exhibit positive

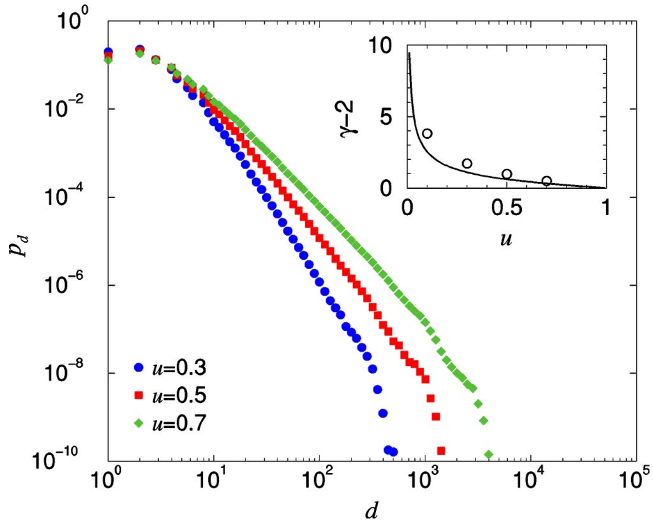

FIG. 8. Degree distribution of the connecting nearest-neighbor model for different values of the addition rate u, graph size $N = 10^6$, and average over 100 realizations. The inset shows the exponent γ obtained from the fit to the power law $p_d = a d^{-\gamma}$ (circles) together with the analytical prediction (continuous line).

degree correlations. This result is in very good agreement with the observations made for social graphs that are also characterized by positive degree correlations. Hence, the connecting nearest-neighbor mechanism generates many of the topological properties of social networks, including power-law degree distributions and positive correlations.

V. DUPLICATION DIVERGENCE

The evolution of some real graphs is given by a replication or partial replication of its local structure. An example is the genome that evolves, among other mechanisms, through single gene or full genome duplications [82] and mutations that lead to the differentiation of the duplicate genes. The evolution of the genome can be translated into the evolution of the protein-protein interaction network where each vertex represents the protein expressed by a gene. After gene duplication both expressed proteins will have the same interactions. This corresponds to the addition of a new vertex in the network with edges pointing to the neighbors of its ancestor. In addition, positive and negative mutations can be modeled by the creation and loss, respectively, of the edges leading to the divergence of the duplicates [28,50,83]. The duplication mechanism has also been considered in the evolution of other biological networks [84]. Moreover, another example is given by the WWW, where new web pages may be created by making a copy or a partial copy of the hyperlinks present in other web pages [85]. In this case the duplication represents the copying process and the divergence the deletion or addition of hyperlinks in the duplicated pages.

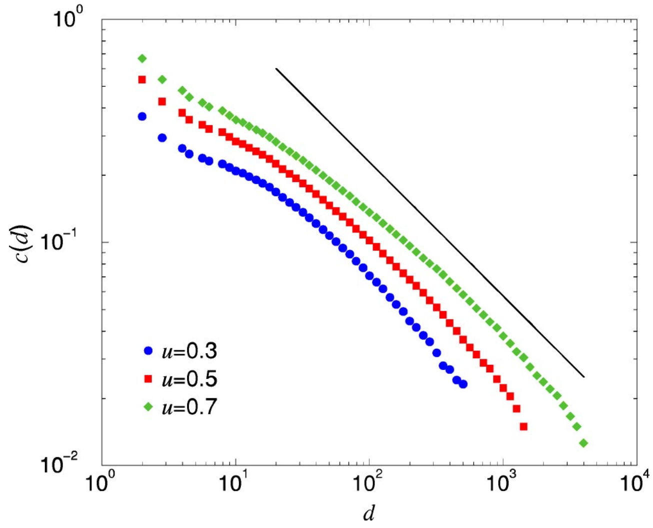

FIG. 9. Clustering coefficient as a function of vertex degree of the connecting nearest-neighbor model for different values of the addition rate u, graph size $N = 10^6$, and average over 100 realizations. The solid line is a power-law decay with exponent 0.6.

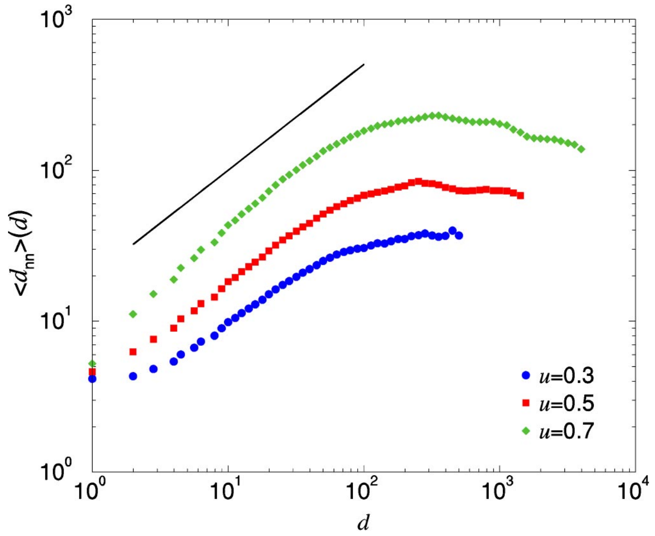

FIG. 10. Average degree among the neighbors of a vertex with degree d of the connecting nearest-neighbor model for different values of the addition rate u, graph size $N = 10^6$, and average over 100 realizations. The solid line is a power-law growth with exponent 0.6.

In a first approximation we will assume that the processes of duplication and divergence are not coupled but take place independently one of the other. Moreover, we will also assume that the creation and deletion of edges take place at random and that they are independent of the degree of the vertices at the edge ends, or any other topological property. Under these approximations, the evolution of the degree of a vertex (the number of interacting partners) is given by

$$\frac{\partial d_i}{\partial N} = \nu_D d_i + \nu_C (N - d_i) - \nu_L d_i , \tag{42}$$

where ν_D, ν_C, and ν_L are the rates per unit of vertex added of duplications, edge creation, and edge lost, respectively. By definition, each duplication implies the addition of a new vertex and, therefore,

$$\nu_D = \frac{1}{N}. \tag{43}$$

We will further assume that

$$\nu_C = \frac{\mu_0}{N}, \quad \nu_L = \frac{\mu_1}{N}; \tag{44}$$

otherwise the stationary graph will be empty or fully connected, both being unreal. Notice that μ_0 and μ_1 are new parameters with no relation to those introduced in the previous section. Then, substituting Eqs. (43) and (44) into Eq. (42) we obtain

$$N\frac{\partial d_i}{\partial N} = \mu_0 + (1-\mu_1) d_i. \tag{45}$$

The linear dependency of the growth rate on d_i evidences once again the existence of an effective linear preferential attachment. The integration of this equation yields

$$d_i(N) = \left(d_i(N_i) + \frac{\mu_0}{1-\mu_1} \right) \left(\frac{N}{N_i} \right)^{\beta} - \frac{\mu_0}{1-\mu_1}, \tag{46}$$

where N_i and $d_i(N_i)$ are the graph size and degree of vertex i when vertex i was added to the graph, and

$$\beta = 1 - \mu_1. \tag{47}$$

Here we have implicitly assumed that

$$\mu_1 < 1; \tag{48}$$

otherwise the stationary state will be an empty graph.

From Eq. (46) it follows that

$$P(d_i > d) = P\left[\left(d_i(N_i) + \frac{\mu_0}{1-\mu_1} \right) \left(\frac{N}{N_i} \right)^{\beta} - \frac{\mu_0}{1-\mu_1} > d \right]. \tag{49}$$

This probability should be computed taking into account that both N_i and $d_i(N_i)$ are random variables. If the duplications take place at a constant rate then the probability density of N_i is given by $P(N_i = N) = 1/N$. Moreover, the probability that a vertex has degree $d_i(N_i)$ when it is introduced is just the probability that its ancestor has this degree. If the graph is in a stationary state then $P[d_i(N_i) = d] = p_d$ is just the degree distribution. Hence

$$P(d_i > d) = \sum_{d'} p_{d'} \int_1^N \frac{dN_i}{N} \Theta\left[\left(d' + \frac{\mu_0}{1-\mu_1} \right) \left(\frac{N}{N_i} \right)^{\beta} - \frac{\mu_0}{1-\mu_1} > d \right]. \tag{50}$$

For $N \gg 1$ we finally obtain

$$p_d = \frac{\partial P(d_i > d)}{\partial d} \sim \left(\frac{\mu_0}{1-\mu_1} + d \right)^{-\gamma} \tag{51}$$

with

$$\gamma = 1 + \frac{1}{1-\mu_1}. \tag{52}$$

The origin of this power-law degree distribution is determined by the second term in the right-hand side of Eq. (45), associated with the vertex duplications and subsequent edge lost. These are local mechanisms and, as in the models describe before, they lead to an effective preferential attachment manifested as a power-law degree distribution.

The next step is thus to investigate if the duplication-divergence model satisfies the inverse proportionality between the average clustering coefficient and vertex degree. If the creation of new interactions takes place at random, i.e., they appear between randomly chosen vertices, then the average clustering coefficient will be negligible for large graph sizes N. There is, however, one source of new interactions giving an appreciable contribution. In the duplication process, if the ancestor is a self-interacting protein then the ancestor and the duplicate may have an interaction among them [28]. Let us assume that this happens with a probability q_v. Thus, if a neighbor of a vertex i is duplicated it will gain a new neighbor (the copy) and with probability q_v an edge between its neighbors (that between the copy and its ancestor), and therefore

$$\frac{\partial e_i}{\partial t} \approx q_v \frac{\partial d_i}{\partial t}, \tag{53}$$

where we have neglected any other process leading to new interactions and edges lost. The integration of this equation yields

$$\langle c \rangle_d = \frac{2e(d)}{d(d-1)} \approx \frac{2q_v}{d}. \tag{54}$$

Hence, under these assumptions we obtain the inverse proportionality behavior. The inclusion of the edge lost may change this result. We do not have any analytical proof but since this process contributes to the loss of triangles and it has a higher impact in high-degree vertices, then we expect that $\langle c \rangle_d$ would decay faster than d^{-1}.

A. Coupled duplication-divergence model

In some practical cases the processes of duplication and divergence cannot be decoupled. For instance, the protein-protein interaction network has a functional role in the organism and, therefore, the lost of certain interactions can result in the death of the corresponding organism. According to the classical model [82] after duplication the duplicate genes have fully overlapping functions. Later on, one of the copies may either become nonfunctional due to degenerative mutations or it can acquire a novel beneficial function and become preserved by natural selection. In a more recent framework [86,87], it is proposed that both duplicate genes are subject to degenerative mutations, losing some functions but jointly retaining the full set of functions present in the ancestral gene. To investigate the influence of the coupling between duplication and divergence we consider the follow-

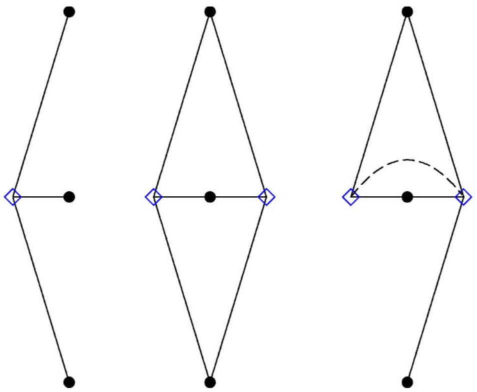

FIG. 11. Schematic representation of the coupled duplication-divergence model evolution rules. Left and middle: a vertex (◇) is being duplicated. Right: the divergence of the duplicates is manifested as a coupled lost of interactions, where the coupling is given by the restriction that for each neighbor (●) at least one of the duplicates should preserve an edge to it. Moreover, due to the existence of self-interactions, a new edge can be created between the duplicates (dashed line).

ing model introduced in Ref. [50]. At each time step a vertex is added according to the following rules.

Duplication. A vertex i is selected at random. A new vertex i' with an edge to all the neighbors of i is created. With probability q_v an edge between i and i' is established (self-interacting proteins).

Divergence. For each of the vertices j connected to i and i' we choose randomly one of the two edges (i,j) or (i',j) and remove it with probability $1-q_e$.

A schematic representation of these rules is shown in Fig. 11. A similar model with an asymmetric divergence was introduced in Ref. [83]. For practical purposes the algorithm starts with two connected vertices and we repeat the duplication-divergence rules N times. Since genome evolution analysis [28,88] supports the idea that the divergence of duplicate genes takes place shortly after the duplication, we can assume that the divergence process always occurs before any new duplication takes place; i.e., there is a time scale separation between duplication and mutation rates. This allows us to consider the number of vertices in the network, N, as a measure of time (in arbitrary units). It is worth remarking that the algorithm does not include the creation of new edges, i.e., the developing of new interactions between gene products, other than those due to self-interactions. However, we have tested that the introduction in the coupled duplication-divergence algorithm of a probability to develop new random connections does not change the network topology substantially.

In order to provide a general analytical understanding of the model, we use a mean-field approach for the moment distribution behavior. Let $\langle d\rangle(N)$ be the average degree of the network with N vertices. After a duplication event $N\rightarrow N+1$ we have that the average degree is given by

$$\langle d\rangle(N+1)=\frac{N\langle d\rangle(N)+2q_v+(2q_e-1)\langle d\rangle(N)}{N+1}. \quad (55)$$

On average, the gain will be proportional to $2q_v$ because of the interaction between duplicates, and to $2\langle d\rangle(N)$ because of duplication, and a loss proportional to $2(1-q_e)\langle d\rangle(N)$ due to the divergence process. For large N, taking the continuum limit, we obtain a differential equation for $\langle d\rangle$. For $q_e<1/2$, $\langle d\rangle$ grows with N but saturates to the stationary value $\langle d\rangle=2q_v/(1-2q_e)+\mathcal{O}(N^{2q_e-1})$. On the contrary, for $q_e>1/2$, $\langle d\rangle$ grows with N as N^{2q_e-1}. At $q_e=q_1=1/2$ there is a dramatic change of behavior in the large scale degree properties. Analogous equations can be written for higher-order moments $\langle d^l\rangle$. Using a rate equations approach similar to that considered in Ref. [89] it is obtained that

$$\frac{\partial n_d}{\partial N}=A_{d-1}n_{d-1}-A_d n_d-\frac{n_d}{N}+2q_v G_{d-1}+2(1-q_v)G_d, \quad (56)$$

where

$$A(d^{in})=\frac{1}{N}(q_v+q_e d), \quad (57)$$

$$G_d=\sum_{d'\geqslant d}\binom{d'}{d}\frac{n_{d'}}{N}\left(\frac{q_e}{2}\right)^d\left(1-\frac{q_e}{2}\right)^{d'-d}. \quad (58)$$

The first two terms in the right-hand side of Eq. (56) result from the duplication of a neighbor of a vertex (with probability $q_e d/N$) and the duplication of a vertex with the creation of an edge between the duplicates (with probability q_v/N), yielding the attachment rate in Eq. (57). Moreover, the last three terms are given by the divergence of the duplicates, where with probability n_d/N a vertex with degree d is replaced by two duplicates (factor of 2 in the last two terms). Thus, the coupling of the duplication and divergence mixes the equations for different n_d. We cannot give an exact derivation of n_d but we can compute the moments of the degree distribution [50,89]. Multiplying Eq. (56) by d^l and summing over d we obtain

$$M_l=\sum_d p_d d^l\sim N^{\sigma_l(q_e)}, \quad (59)$$

where

$$\sigma_l(q_e)=lq_e+2\left[\left(\frac{1+q_l}{2}\right)^l-1\right], \quad (60)$$

provided $\sigma_l(q_e)>0$. If $\sigma_l(q_e)<0$ the corresponding moment approaches a stationary value for large N. For all l we find a value q_l at which the moments cross from a divergent behavior to a finite value for $N\rightarrow\infty$. In particular, for $l=1$ we have $q_1=1/2$ (as obtained above) and for $l=2$ we obtain $q_2=2\sqrt{3}-3\approx 0.46$. Moreover, the nonlinear behavior with l is indicative of a multifractal degree distribution.

In order to support the analytical calculations, we have performed numerical simulations of the coupled duplication-divergence model with graph size ranging from $N=10^3$ to 10^6. In Fig. 12 we report the generalized exponents $\sigma_l(q_e)$ as a function of the divergence parameter q_e. As predicted by the analytical calculations, $\sigma_l=0$ at a critical value q_l.

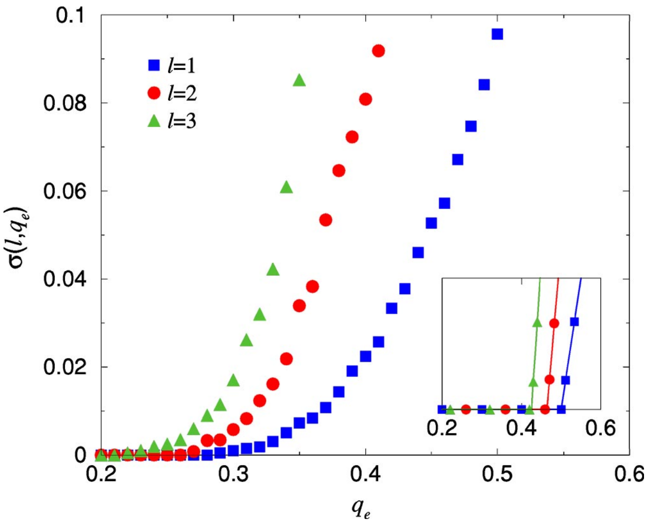

FIG. 12. The exponent $\sigma_l(q_e)$ as a function of q_e for different values of l. The symbols were obtained from numerical simulations of the model. The moments $\langle d^l\rangle$ were computed as a function of N in networks with size ranging from $N=10^3$ to $N=10^6$. The exponents $\sigma_l(q)$ are obtained from the power-law fit of the plot $\langle d^l\rangle$ vs N. In the inset we show the corresponding mean-field behavior, as obtained from Eq. (60), which is in qualitative agreement with the numerical results.

The general phase diagrams obtained is in good qualitative, but not quantitative, agreement with the mean-field predictions and the multifractal picture. Noticeably, multifractal features are present also in a recently introduced model of growing networks [49] where, in analogy with the duplication process, newly added vertices inherit the network degree properties from parent vertices. Multifractality thus appears to be related to local inheritance mechanisms. Multifractal distributions have a rich scaling structure where the scale-free behavior is characterized by a continuum of exponents. This behavior is, however, opposite to that of the usual exponentially bounded distributions. Even if the evolution rules of the coupled duplication-divergence model are local they introduce an effective linear preferential attachment. However, because the edge deletion of duplicate vertices introduce additional heterogeneity in the problem, we obtain a multifractal behavior.

The coupling between duplication and divergence is however less relevant to determine the scaling of the average clustering coefficient with vertex degree. In fact, for the coupled duplication-divergence model Eq. (53) also applies, obtaining the inverse proportionality in Eq. (54). In Fig. 13 we plot $\langle c\rangle_d$ vs d for different values of q_e, manifesting a power-law decay but with an exponent larger than 1. With decreasing q_e (increasing the loss of edges) the power-law decay deviates more and more from the predicted behavior $\langle c\rangle_d \sim d^{-1}$. This picture corroborates our hypothesis that if the edge loss is sufficiently large then a faster decay should be observed.

On the other hand, the average neighbor degree as a function of the vertex degree for different values of q_e is depicted in Fig. 14. Negative degree correlations are manifested by a power-law decay $\langle d_{nn}\rangle \sim d^{-0.1}$. The existence of negative degree correlations has been actually reported in Ref. [90] for a protein-protein interaction network. Moreover, a model based on these correlations has also been proposed in Ref. [91].

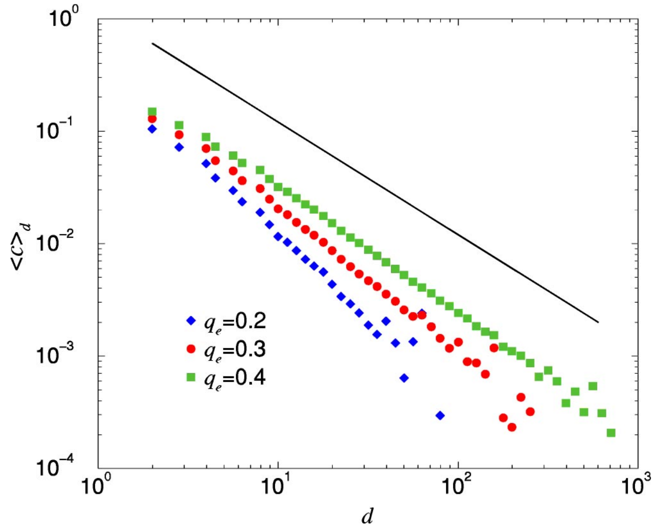

FIG. 13. Clustering coefficient as a function of vertex degree of the coupled duplication-divergence model for different values of q_e, graph size $N=10^6$, and average over 100 realizations. The solid line is a power-law decay with exponent 1.

VI. DISCUSSION AND CONCLUSIONS

After analyzing these models we can conclude that growing networks based on local evolution rules exhibit an effective linear preferential attachment. The general principle behind it is the following. It is true that when we take a vertex at random the selection does not imply any degree preference, other than the one imposed by the degree distribution. However, if we take a neighbor of that vertex then some preference is induced. In fact, the probability that vertex i is

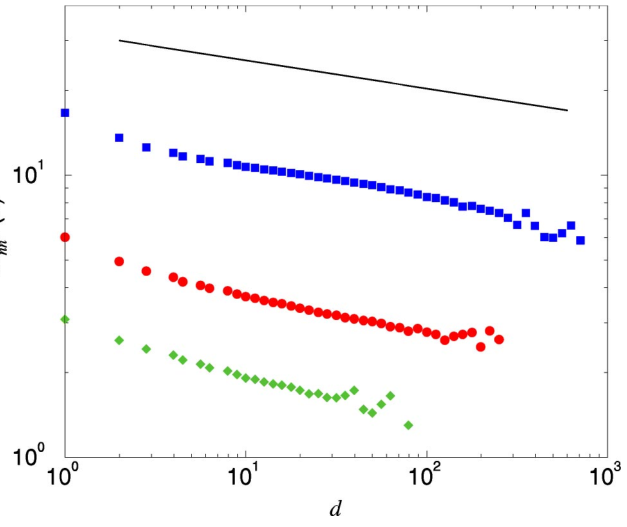

FIG. 14. Average degree among the neighbors of a vertex with degree d of the coupled duplication-divergence model for different values of q_e, graph size $N=10^6$, and average over 100 realizations. The solid line is a power-law decay with exponent 0.1.

TABLE I. Summary of the correlation properties of the different models analyzed here.

Mechanism	$\langle c\rangle_d \sim d^{-\beta}$	$\langle d_{nn}\rangle_d \sim d^{\alpha}$
Connecting neighbors	$0<\beta<1$	$\alpha>0$
Random walk	$\beta=1$	$\alpha\leqslant 0$
Duplication divergence	$\beta\geqslant 1$	$\alpha<0$

a neighbor of the randomly selected vertex is simply

$$\frac{d_i}{\sum_j d_j}, \tag{61}$$

which is exactly the linear preferential attachment considered in the BA model [19]. Therefore, the connection to a neighbor of a vertex selected at random leads to an effective linear preferential attachment.

Another important consequence of the local models considered above is the inverse proportionality between the average clustering coefficient and the vertex degree, or more generally $\langle c\rangle_d \sim d^{-\beta}$. This result is determined by the fact that when a new edge is created to a vertex then with a certain probability an edge will also be created to one or more of its neighbors. Thus, locality is again a crucial point. On the other hand, even if we were not able to find an analytical explanation, these local models are also characterized by degree correlations among connected vertices.

These features are observed in the three models analyzed here and are summarized in Table I. They describe different systems such as technological, social, and biological networks, which appear unrelated from the definitions of their evolution rules. The detailed analysis performed here reveals that their main property that they are local models of growing networks, explains the existence of strong similarities in their topological properties. These observations can be extended to other local models proposed in the literature. An example is the model introduced in Ref. [92], where each time a vertex is added it is connected to both ends of an edge selected at random. It can be easily shown that this rule also introduces an effective linear preferential attachment, clustering hierarchy, and degree correlations. Another example is the deactivation model [60], where new vertices are connected to small subset of connected vertices. A detailed study of its topology [63] reveals the existence of clustering hierarchy and degree correlations.

In conclusion, the growing models with local rules exhibit some of the common features of real graphs. They are characterized by an effective preferential attachment, an average clustering coefficient that decreases with increasing vertex degree, and degree correlations. The local knowledge is then a general principle determining the topology of growing complex networks.

ACKNOWLEDGMENTS

I thank A. Vespignani, Y. Moreno, and A.-L. Barabási for helpful comments and discussion.

[1] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[2] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University Press, Princeton, NJ, 1999).

[3] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2001).

[4] S. N. Dorogovtsev and J. F. F. Mendes., Adv. Phys. **51**, 1079 (2002).

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

[6] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).

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

[8] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, e-print cond-mat/0206084.

[9] Z.E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).

[10] The National Laboratory for Applied Network Research (NLANR), National Science Foundation, Washington, DC, http://moat.nlanr.net/

[11] Internet mapping project, Lucent Bell Labs, Murray Hill, NJ, http://www.cs.bell-labs.com/who/ches/map/

[12] Information Sciences Institute, http://www.isi.edu/div7/scan/

[13] The Cooperative Association for Internet Data Analysis (CAIDA), San Diego Supercomputer Center, San Diego, CA, http://www.caida.org/home/

[14] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).

[15] R. Govindan and H. Tangmunarunkit, in *Proceedings of the 2000 IEEE INFOCOM Conference, Tel Aviv, Israel, 2000* (IEEE Service Center, Piscataway, NJ, 2000), pp. 1371–1380.

[16] G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).

[17] S.-H. Yook, H. Jeong, and A.-L. Barabási, Proc. Natl. Acad. Sci. U.S.A. **99**, 13382 (2002).

[18] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).

[19] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi, Science **287**, 2115a (2000).

[20] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Comput. Netw. **33**, 309 (2000).

[21] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103(R) (2002)

[22] M. E. J. Newman, S. Forrest, and J. Balthrop, Phys. Rev. E **66**, R035101 (2002).

[23] R. F. Cancho, C. Janssen, and R. V. Sol, Phys. Rev. E **64**, 046119 (2001).

[24] P. Uetz *et al.*, Nature (London) **403**, 623 (2000).

[25] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, Proc. Natl. Acad. Sci. U.S.A. **97**, 1143 (2000).

[26] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, and Y. Sakaki, Proc. Natl. Acad. Sci. U.S.A. **98**, 4569 (2001).

[27] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).

[28] A. Wagner, Mol. Biol. Evol. **18**, 1283 (2001).

[29] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000).

[30] A. Wagner and D. Fell, Proc. R. Soc. London, Ser. B **268**, 1803 (2001).

[31] J. Camacho, R. Guimerá, and L. A. N. Amaral, Phys. Rev. Lett. **88**, 228102 (2002).

[32] R. V. Solé and J. M. Montoya, Proc. R. Soc. London, Ser. B **268**, 2039 (2001).

[33] J. Lahererre and D. Sornette, Eur. Phys. J. B **2**, 525 (1998).

[34] S. Redner, Eur. Phys. J. B **4**, 131 (1998).

[35] A. Vázquez, e-print cond-mat/0105031.

[36] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).

[37] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).

[38] E. M. Jin, M. Girvan, and M. E. J. Newman, Phys. Rev. E **64**, 046132 (2001).

[39] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A **311**, 590 (2002).

[40] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Berg, Nature (London) **411**, 907 (2001).

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

[42] A.-L. Barábasi, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).

[43] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).

[44] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).

[45] P. L. Krapivsky and S. Redner, Phys. Rev. E **63**, 066123 (2001).

[46] S. N. Dorogovtsev and J. F. F. Mendes, Phys. Rev. E **62**, 1842 (2000).

[47] S. Mossa, M. Barthélémy, H. E. Stanley, and L. A. N. Amaral, Phys. Rev. Lett. **88**, 138701 (2002).

[48] G. Bianconi and A.-L. Barabási, Europhys. Lett. **54**, 436 (2000).

[49] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Europhys. Lett. **57**, 334 (2002).

[50] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, ComPlexUs **1**, 38 (2003).

[51] H. Jeong, Z. Néda, and A.-L. Barabási, e-print cond-mat/0104131.

[52] M. E. J. Newman, Phys. Rev. E **64**, R025102 (2001).

[53] B. Bollobás, in *School on Statistical Physics, Probability Theory and Computational Complexity* (International Centre for Theoretical Phyics, Trieste, 2002).

[54] M. Boguña and R. Pastor-Satorras, Phys. Rev. E **66**, 047104 (2002).

[55] J. Berg and M. Lassig, Phys. Rev. Lett. **89**, 228701 (2002).

[56] A. Vázquez and M. Weigt, Phys. Rev. E **67**, 027101 (2003).

[57] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001); M. Boguña, R. Pastor-Satorras, and A. Vespignani, *ibid.* **90**, 028701 (2003).

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

[59] D. S. Callaway, J. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz, Phys. Rev. E **64**, 041902 (2001).

[60] K. Klemm and V. M. Eguíluz, Phys. Rev. E **65**, 036123 (2002).

[61] P. Holme and B. J. Kim, Phys. Rev. E **65**, 026107 (2002).

[62] G. Szabó, M. Alava, and J. Kertész, Phys. Rev. E **66**, 026101 (2002).

[63] A. Vázquez, M. Boguña, Y. Moreno, R. Pastor-Satorras, and A. Vespignani, e-print cond-mat/0207711.

[64] A. Capocci, G. Caldarelli, R. Marchetti, and L. Pietronero Phys. Rev. E **64**, 035105 (2001).

[65] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz Phys. Rev. Lett. **89**, 258702 (2002).

[66] K.-I. Goh, B. Kahng, and D. Kim Phys. Rev. Lett. **88**, 108701 (2002).

[67] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2002).

[68] P. L. Krapivsky and S. Redner, Comput. Netw. **39**, 261 (2002).

[69] *GVU's Ninth WWW User Survey Report* (Georgia Tech Research Corporation, Atlanta, 1998).

[70] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, in *Proceedings of the 8th International World Wide Web Conference, Toronto, Canada, 1999*, edited by A. Mendelzon (Elsevier Science, New York, 1999) pp. 213–225.

[71] S. Brin and L. Page, Comput. Netw. **30**, 107 (1998).

[72] A. Vázquez, Europhys. Lett. **54**, 430 (2001).

[73] B. Tádic and D. Dhar, Phys. Rev. Lett. **79**, 1519 (1997).

[74] A. Vázquez and O. Sotolongo-Costa, J. Phys. A **32**, 2633 (1999).

[75] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. Lett. **59**, 381 (1987).

[76] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. A **38**, 364 (1988).

[77] A. Vespignani and S. Zapperi, Phys. Rev. Lett. **78**, 4793 (1997).

[78] A. Vespignani and S. Zapperi, Phys. Rev. E **57**, 6345 (1998).

[79] J. Davidsen, H. Ebel, and S. Bornholdt, Phys. Rev. Lett. **88**, 128701 (2001).

[80] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **281**, 69 (2000).

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

[82] S. Ohono, *Evolution by Gene Duplication* (Springer-Verlag, Berlin, 1970).

[83] R. V. Solé, R. Pastor-Satorras, E. D. Smith, and T. Kepler, Adv. Complex Syst. **5**, 43 (2002).

[84] F. Slanina and M. Kotrla, Phys. Rev. E **62**, 6170 (2000).

[85] J. Kleinberg, R. Kumar, P. Raphavan, S. Rajagopalan, and A. Tomkins, in *Proceedings of the 5th International Conference on Combinatorics and Computing, Tokyo, Japan, 1999*, edited by T. Asano, H. Imai, D. T. Lee, S. Nakano, and T. Tokuyama (Springer-Verlag, Heidelberg, 1999), pp. 1–17.

[86] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. l. Yan, and J. Postlethwait, Genetics **151**, 1531 (1999).

[87] M. Lynch and A. Force, Genetics **154**, 459 (1999).

[88] M. A. Huynen and P. Bork, Proc. Natl. Acad. Sci. U.S.A. **95**, 5849 (1998).

[89] J. Kim, P. L. Krapivsky, B. Kahng, and S. Redner, Phys. Rev. E **66**, 055101 (2002).

[90] S. Maslov and K. Sneppen, Science **296**, 910 (2002).

[91] J. Berg, M. Lässig, and A. Wagner, e-print cond-mat/0207711.

[92] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **63**, 062101 (2001).

[93] http://www.nd.edu/126networks