Self-organization of collaboration networks

José J. Ramasco,[1,*] S. N. Dorogovtsev,[2,3,4] and Romualdo Pastor-Satorras[5]
[1]*Departamento de Física and Centro de Física do Porto, Faculdade de Ciências, Universidade do Parto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal*
[2]*Departmento de Física, Universidade de Aveiro, Campus Universitario de Santiago, 3810-193 Aveiro, Portugal*
[3]*Laboratory of Physics, Helsinki University of Technology, FIN-02015 HUT, Finland*
[4]*A. F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia*
[5]*Departament de Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, Campus Nord, 08034 Barcelona, Spain*
(Received 29 March 2004; published 14 September 2004)

We study collaboration networks in terms of evolving, self-organizing bipartite graph models. We propose a model of a growing network, which combines preferential edge attachment with the bipartite structure, generic for collaboration networks. The model depends exclusively on basic properties of the network, such as the total number of collaborators and acts of collaboration, the mean size of collaborations, etc. The simplest model defined within this framework already allows us to describe many of the main topological characteristics (degree distribution, clustering coefficient, etc.) of one-mode projections of several real collaboration networks, without parameter fitting. We explain the observed dependence of the local clustering on degree and the degree–degree correlations in terms of the "aging" of collaborators and their physical impossibility to participate in an unlimited number of collaborations.

DOI: 10.1103/PhysRevE.70.036106 PACS number(s): 89.75.Hc, 87.23.Ge, 05.70.Ln

I. INTRODUCTION

Recent years have witnessed an upsurge in the study of complex systems that can be described in terms of networks, in which the vertices picture the elementary units composing the system, and the edges represent the interactions or relations between pairs of units [1,2]. These studies have led to the development of a modern theory of complex networks which has found fruitful applications in fields as diverse as the Internet [3], the World Wide Web [4], or biological interacting networks [5–8].

An important example of this kind of system, that has attracted a great deal of interest from researchers in different scientific fields, is social networks [9]. The study of social networks has been traditionally hindered by the small size of the networks considered and the difficulties in the process of data collection (usually from questionnaires or interviews). More recently, however, the increasing availability of large digital databases has allowed one to study a particular class of social networks, the so-called *collaboration networks*. These networks can be defined in a nonambiguous way, and their exceptionally large size has permitted empirical researchers to obtain a reliable statistical description of their topological properties and to arrive at solid conclusions concerning their structure.

Social collaboration networks are generally defined in terms of a set of people (called *actors* in the social science literature) and a set of *collaboration acts*. Actors relate to each other by the fact of having participated in a common collaboration act. Examples of this kind of network can be found in movie actors related by costarring the same movie, scientists related by coauthoring a scientific paper, members of the boards of company directors related by sitting on the same board, etc. Collaboration networks can be represented as bipartite graphs [10] with two types of vertices, one kind representing the actors, while vertices of the other kind are acts of collaboration. As a rule, however, it is the one-mode projections of these bipartite graphs that are empirically studied. In these projections, the vertices representing the acts of collaboration are excluded, and collaborating pairs of actors are connected by edges. Since multiple connections in the projected graph are usually ignored, the projection is less informative than the original bipartite graph.

The study of several examples of large collaboration networks [11–15] allows one to draw a number of conclusions regarding the main topological properties of one-mode projections of these networks.

(1) The degree distribution $P(k)$, defined as the probability that a vertex is connected to k other vertices, often exhibits a fat tail that can be approximated by a power law behavior for large k.

(2) The clustering coefficient, roughly defined as the probability that two neighbors of any given vertex are also neighbors of each other, takes in average large values, and it locally depends on the vertex degree, signaling the presence of a structure in the network [16,17].

(3) The degrees of the nearest-neighbor vertices are positively correlated, i.e., vertices with large degree have a high probability to be connected to vertices with large degree, and vice versa. This property has been dubbed *assortative mixing* [18].

The general presence of these three properties in most collaboration graphs prompts toward the development of models capable of reproducing and explaining these features. In general, the first insight into the architecture of a complex network is provided by "formal" constructions of random graphs. These constructions allow one to reproduce the struc-

*Electronic address: jjramasc@fc.up.pt

ture of complex networks, but completely ignore the mechanisms underlying these architectures. The minimal formal model of a complex one-partite graph, that is a graph composed by a single type of vertices, is the configuration model [19–22]. In simple terms, the configuration model generates (uncorrelated) graphs, which are maximally random under the constraint that their degree distribution is a given one. Similarly, the minimal model of a complex bipartite graph is a bipartite network that is maximally random under the constraint that the two degree distributions for both kinds of vertices are given [23,24]. One can see that this is a direct generalization of the configuration model to bipartite graphs. The quality of the configuration model applied to bipartite graphs was checked in Ref. [23]. In this work, it was proved that the empirical degree distribution of the one-mode projection of a bipartite collaboration graph agrees with that of the configuration model when the empirically observed degree distributions are imposed on the two kinds of vertices. One should emphasize that a one-mode projection of an uncorrelated bipartite graph is correlated. In particular, this projection contains numerous triangles of edges which results in a high clustering [25].

Therefore, it might seem at first sight that, in order to explain the nature of the structure of collaboration networks, it is sufficient (i) to propose a mechanism generating the specific degree distributions of the two kinds of vertices and afterwards (ii) to connect vertices by using the configuration model. This approach, however, fails to reproduce the complex distribution of connections over collaboration networks, since it assumes pure randomness. Also, it does not explain specific distributions of vertex degrees, which, in the configuration model, are assumed to be given. Note that, while providing reasonable values of clustering, the configuration model fails to reproduce the type of degree–degree correlations in collaboration networks. Consequently, in order to fully explain the specific architecture of collaboration networks (fat-tailed degree distributions, high clustering, assortative mixing, etc.), we have to introduce a mechanism for the linking of vertices in these networks.

In the present paper we propose a first approximation to such a mechanism. In our approach, we treat collaboration networks as growing, self-organizing, correlated bipartite graphs, applying the ideas at the basis of the preferential attachment concept put forward by Barabási and Albert [26] in the network modeling context (see also Ref. [27]) to bipartite graphs. The simplest model that we can define already allows us to quantitatively describe most of the empirical data on collaboration networks without fitting, only by using basic numbers characterizing the real networks. We emphasize that the absence of fitting convincingly proves the validity of the concept.

The degree–degree correlations in the one-mode projections of collaboration graphs are a topic of our special interest. We show that the "assortative mixing" character of these correlations is not so inevitable in collaboration networks, as it is usually believed [25]. We explain the origin of the assortative mixing in real collaboration networks in terms of the aging of actors, which cannot accept new connections during the whole growth process of the network.

The present paper is organized as follows. In Sec. II we review measurements defined to characterize the topological properties of collaboration networks:—bipartite graphs and their one-mode projections. Section III presents the existing empirical data on collaboration networks, referring in particular to the networks of movie actors, scientific coauthorship, and company directors. In Sec. IV we introduce a simple model of a growing, self-organizing bipartite graph. Section V contains results obtained for this model and a detailed comparison with empirical data. Separately, in Sec. VI we discuss and explain the presence of positive correlations between the degrees of the nearest-neighbor vertices in collaboration networks. In this section we discuss the importance of the "physical" limitation of vertex degrees in collaboration networks. Finally, in Sec. VII we draw the main conclusions of our work.

II. STRUCTURAL ORGANIZATION OF COLLABORATION NETWORKS

As we have already mentioned in the Introduction, collaboration networks can be represented as bipartite graphs [28]. On one side, we have collaboration acts (e.g., movie costarring or paper coauthorship, belonging to the same company, school, etc.) that may be represented as a special kind of vertices. On the other side we have the actors (normal vertices) that are linked to the collaboration acts in which they participate. Two independent degree distributions may then be defined: First, the probability $S(n)$ of having n actors participating in any collaboration act; and second, the probability $Q(q)$ that any actor has taken part in q collaboration acts.

In most of cases, however, the object of study is not the whole bipartite graph but its one-mode projection: i.e., the network formed by the collaborating actors linked to each other whenever they have shared a collaboration act. For this projected network, another degree distribution $P(k)$ may be considered, defined as the probability that any given actor is connected to k others. Focusing on the one-mode projection of a collaboration network, many other properties generally studied in common random graphs can be measured. This type of study has already been carried out for several empirical social networks (see Ref. [29] for a recent review). The quantities that we use to describe the structure of the projected network are the clustering coefficient and the mean clustering [12], the average clustering coefficient of vertices of degree k [30–33], the average degree of the nearest neighbors of the vertices of degree k [34], and the Pearson correlation coefficient defined in Refs. [18,35].

The local clustering c_i of the vertex i is given by the rate between the number of triangles connected to that vertex, s_i, and the total number of possible triangles including it, $k_i(k_i-1)/2$, i.e.,

$$c_i = \frac{2s_i}{k_i(k_i-1)}. \tag{1}$$

To obtain the mean degree-dependent local clustering we average the local clustering over all vertices with degree k in a network,

$$c(k) = \frac{s(k)}{k(k-1)/2}, \tag{2}$$

where $s(k) = \langle s_i(k) \rangle$ is the mean number of connections between the nearest neighbors of a vertex of degree k. The mean clustering $\langle c \rangle$ is defined as the average of the local clustering over all the vertices in a network, i.e.,

$$\langle c \rangle = \sum_{k>1} P(k)c(k) = \frac{1}{N}\sum_i c_i, \tag{3}$$

where N is the total number of actors (vertices), and the second sum runs over the N vertices of the network. The clustering coefficient of a graph (transitivity in sociology [9]) is defined as

$$c = \frac{3 \times (\text{number of triangles of edges in a graph})}{(\text{number of connected triples of vertices})} = \frac{2\sum_k P(k)s(k)}{\sum_k P(k)k(k-1)}. \tag{4}$$

The quantities $c(k)$, $\langle c \rangle$, and c provide information on the concentration of loops of length three in a graph, which is typically high in social networks [36]. Note that if the local clustering depends on the degree, $c \neq \langle c \rangle$, and the (relative) difference is great in many real-world networks.

The correlations between the degrees of connected vertices can be fully defined by means of the joint probability $P(k,k')$, defined such that $(2-\delta_{k,k'})P(k,k')$ is the probability that a randomly chosen edge connects to vertices of degree k and k' [37]. ($\delta_{k,k'}$ is the Kronecker symbol.) By using this quantity one can compute the average degree of the nearest neighbors of the vertices of degree k, $\overline{k}_{nn}(k)$, defined as

$$\overline{k}_{nn}(k) = \langle k \rangle \frac{\sum_{k'} k' P(k,k')}{kP(k)} \equiv \sum_{k'} k' P(k'|k), \tag{5}$$

where $P(k'|k)$ is the conventional probability that a vertex of degree k is connected to a vertex of degree k'. In simple terms, if the network presents assortative mixing (large degree vertices connect preferably with large degree vertices, and vice versa), $\overline{k}_{nn}(k)$ increases with k [38]. In the case of disassortative mixing (large degree vertices connected with low degree vertices, and vice versa), $\overline{k}_{nn}(k)$ is conversely a decreasing function of k. Analogous information can be obtained by means of the Pearson correlation coefficient, defined as

$$r = \langle k \rangle \frac{\sum_k k^2 \overline{k}_{nn}(k) P(k) - \langle k^2 \rangle^2}{\langle k \rangle \langle k^3 \rangle - \langle k^2 \rangle^2}. \tag{6}$$

Here positive (negative) values of r imply the presence of assortative (disassortative) mixing.

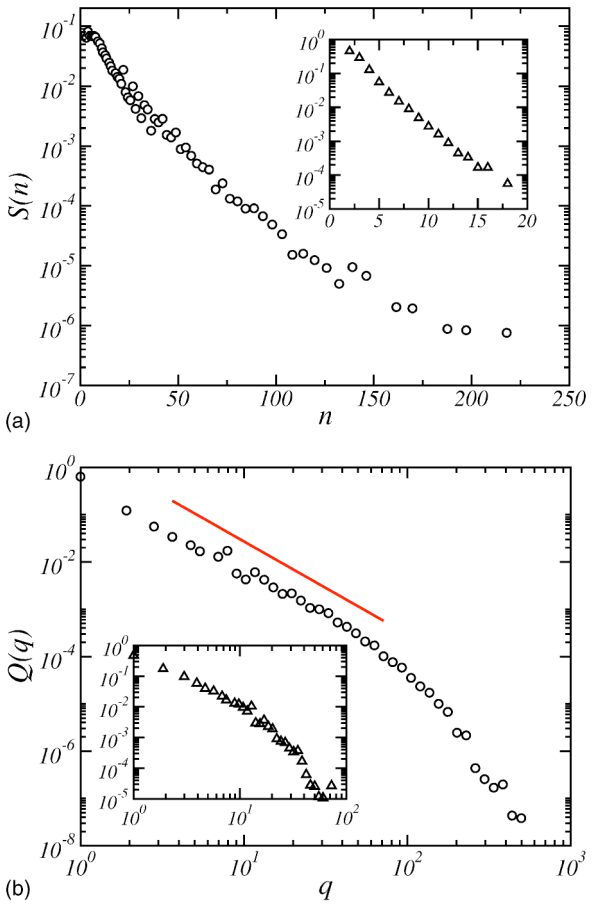

FIG. 1. (a) Probability distribution of the size of collaboration acts $S(n)$ for the movie actor collaboration network (main plot) and the scientific collaboration network (inset). (b) Probability distribution that an actor has taken part in q collaboration acts $Q(q)$ for the movie actor collaboration network (main plot) and the scientific collaboration network (inset). The solid line has slope $\simeq 2$.

III. EMPIRICAL DATA ON COLLABORATION NETWORKS

In the present section we revisit the empirical analysis of three typical social collaboration networks. We consider in particular the network formed by movie actors playing in the same movie, the network of scientific collaborations, and the network of company board directors sitting on the same board.

A. Movie actor collaboration network

The movie actor collaboration network that we consider was obtained from the Internet Movie Database (IMDB) [39]. Taking only into account movies with more than one actor, and discarding duplicated actors in several movies, we finally analyze the properties of a network composed by $N = 382\,219$ actors acting on $t = 118\,477$ films. The distribution of movie cast size, $S(n)$, is represented in Fig. 1(a). Apparently, this function follows an exponential decay, with an average cast size of $\overline{n} = 12.33$ actors per movie. The distribution $Q(q)$ [number of movies in which an actor has played]

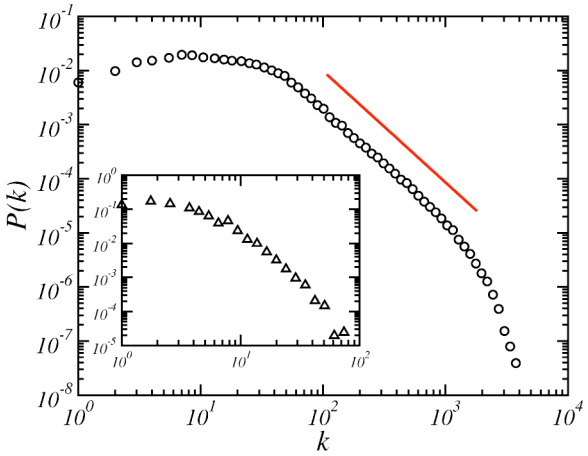

FIG. 2. Degree distribution $P(k)$ of the one-mode projection for the movie actor collaboration network (main plot) and the scientific collaboration network (inset). The full line has slope $\simeq 2$.

adjusts better to a power law decay $Q(q) \sim q^{-\gamma}$ with an apparent exponent $\gamma \approx 2$, see Fig. 1(b). An upper cutoff of this dependence is observed around $q_c \sim 100$. The mean number of movies played per actor is $\langle q \rangle = 3.82$.

The degree distribution of the one-mode projection of this network, $P(k)$, is plotted in Fig. 2. It has a power law decay with approximately the same exponent as $Q(q)$, which extends for close to two decades up to a sharp cutoff at $k_c \sim 2000$. The mean degree of the network is $\langle k \rangle = 78.69$ The local clustering as a function of degree is depicted in Fig. 3. We can observe a flat region, extending up to degree values close to 10^2, followed by a rapid decrease. The mean clustering of the one-mode projection is $\langle c \rangle = 0.78$ and the clustering coefficient is $c = 0.17$. The correlations in the projected network, presented in the form of the average degree of the nearest neighbors of a vertex versus its degree, are plotted in Fig. 4. The increasing behavior of the function $\overline{k}_{nn}(k)$ is compatible with the presence of assortative mixing, a fact that is further confirmed by the value of the Pearson coefficient, $r = 0.23$. In Table I we summarize the main average values obtained for this network.

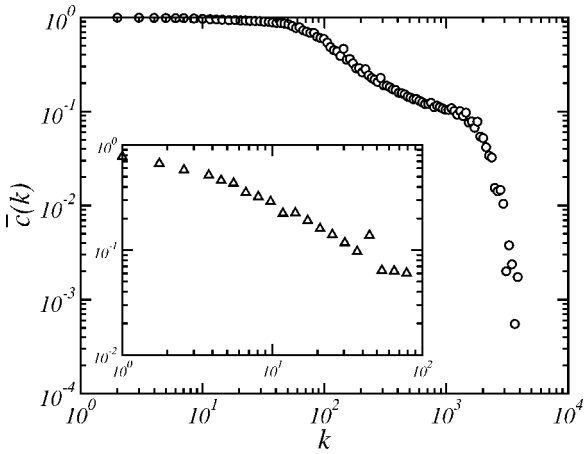

FIG. 3. Local clustering as a function of the degree $c(k)$ for the movie actor collaboration network (main plot) and the scientific collaboration network (inset).

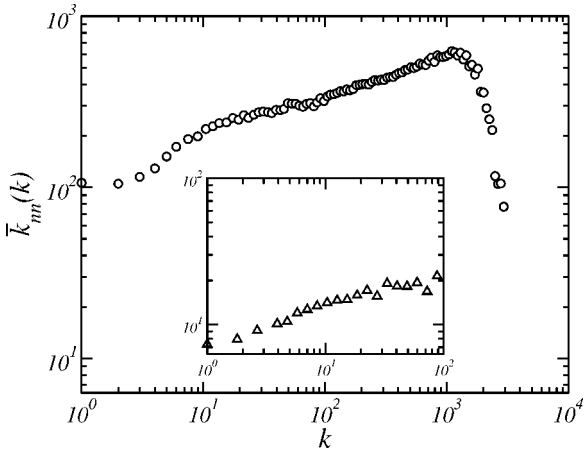

FIG. 4. Average degree of the nearest neighbors as a function of the degree $\overline{k}_{nn}(k)$ for the movie actor collaboration network (main plot) and the scientific collaboration network (inset).

B. Scientific collaboration networks

The next collaboration network that we analyze is the network of scientific collaborations collected from the condensed matter preprint database at Los Alamos [40]. In this collaboration graph, the actors represent scientists which have collaborated in the writing of a scientific paper. The complete bipartite network is composed by $t = 17\,828$ papers and $N = 16\,258$ authors. The distribution of the number of authors in a given paper is plotted in Fig. 1(a). This distribution is clearly exponential, with an average value $\overline{n} = 3.05$. The distribution of the number of papers written by any given author, Fig. 1(b), shows, on the other hand, an apparent power law behavior, even though the limited range that it takes (scarcely more than one decade) precludes the determination of a significant exponent. The average number of papers written by any author is in this case $\langle q \rangle = 3.35$.

The degree distribution of the one-mode projection of the scientific collaboration network, plotted in Fig. 2, shows again a fat-tailed behavior, compatible with a power law. The corresponding average degree is $\langle k \rangle = 5.85$. The degree-dependent local clustering $c(k)$ and the average degree of the nearest neighbors are explored in Figs. 3 and 4, respectively. This last result, together with a Pearson $r = 0.31$, indicates the presence of a strong assortative mixing. Additional numerical parameters characterizing this network are summarized in Table I.

C. Board of directorships

The last collaboration network that we report is the network of company directors, in which two directors are linked if they sit on the same board of directors. Table I reports the data corresponding to the list of the "Fortune 1000" US companies, obtained from Refs. [29,36]. It includes $t = 914$ companies and $N = 7673$ directors. The average number of directors per company is $\overline{n} = 11.5$. Both distributions $Q(q)$ and

TABLE I. Comparison of calculations with empirical data for the movie, scientific collaboration, and codirectorship networks and the simulations of the model.

	Movie actors network	Analytic results	Numeric results	Numeric with aging	Coauthors network	Analytic results	Numeric results	Numeric with aging	Directors network	Analytic results	Numeric results
t	118477		10^5	10^5	17828		10^5	10^5	914		10^5
N	382219				16258				7673		
$\overline{n}$	12.33	12.33	12.33	12.33	3.05	3.05	3.05	3.05	11.5	11.5	11.5
$\overline{m}$		3.23	3.23	3.23		0.91	0.91	0.91		8.39	8.39
q_c	$\sim 10^2$			10^2	~ 15			15	~ 10		
$\langle q \rangle$	3.82	3.82	4.44	4.39	3.35	3.35	3.70	3.69		1.37	1.48
$\langle k \rangle$	78.69	43.25	75.05	85.67	5.85	6.87	8.45	8.93	14.44	14.39	17.10
γ		2.35				2.43				4.70	
$\langle c \rangle$	0.78	0.71	0.76	0.70	0.64	0.68	0.50	0.43	0.88	0.87	0.86
c	0.17	0.06[a]	0.037	0.08	0.36	0.08[b]	0.026	0.09	0.59	0.5	0.32
r	0.23		−0.13	0.14	0.31		−0.08	0.40	0.28		0.11

[a]The cutoff $k_c \sim 10^3$ was used.
[b]The cutoff $k_c \sim 10^2$ was used.

$P(k)$ can be adjusted by exponential decaying functions, although the range of values for q and k is quite restricted. The mean degree of the projected network is $\langle k \rangle = 14.44$. The clustering coefficient of the one-mode projected network is quite large, and it shows a clear assortative mixing behavior, as given by a Pearson correlation coefficient $r = 0.28$.

IV. SELF-ORGANIZED COLLABORATION MODEL

A. Definition of the model

To understand the common properties of collaboration networks, we propose a self-organized growing model. We exploit two generic features of collaboration networks: (i) Social collaboration networks are organized as bipartite graphs. (ii) Social collaboration networks are not static entities, but they grow in time by the continuous addition of new acts of collaboration (movies produced or papers written), and new actors, that increase the pool of possible participants in new acts of collaboration.

Using the language of movies to make the description more concrete, our growing bipartite network model is defined by the following rules:

(1) At each time step a new movie with n actors is added.

(2) Of the n actors playing in a new movie, m actors are new, without previous experience.

(3) The rest $n-m$ actors are chosen from the pool of "old" actors with a probability proportional to the number q of movies that they previously starred.

The total number of movies is t, the "time." The number n may be either constant or a random variable distributed with a given distribution $S(n)$. The number m may also be either constant or a random variable. At each time step, the total number of actors increases as $N \to N+m$. Thus the model generates a bipartite graph of t movie vertices and N actor vertices. Note that the proportional preference corresponds to the following practical rule of selection of actors: A director randomly selects a previous movie and then chooses at random one of its actors.

B. Analytical results

One can see that the evolution rules of the present model practically coincide with those of the Simon model [27]. For simplicity, let us assume that the number of actors playing in each movie is constant and equal to its average value, $n = \overline{n}$, as well as the number of new actors per movie, $m = \overline{m}$. This assumption is in fact quite reasonable given the exponential nature of the $S(n)$ distributions observed empirically. We also assume that if the total number of actors is large, the probability that two actors selected for a new movie have already costarred in other old film is vanishingly small. Note that, strictly speaking, this assumption is only valid for uncorrelated networks with rapidly decreasing degree distributions. Nevertheless, the results obtained with it provide a good enough approximation to the empirical values (see Table I) to justify its introduction.

Within this approximation, since each movie starred by an actor leads to the acquisition of $\overline{n}-1$ new coactors, we have a strict relation between the experience of an actor, q, and the total number of its coactors (its degree in the projected network), k:

$$k = q(\overline{n} - 1). \tag{7}$$

In particular, at large k and q, when we can consider both variables to be continuous, we have

$$P(k) \cong \frac{1}{\overline{n}-1} Q\left(\frac{k}{\overline{n}-1}\right). \tag{8}$$

In the limit of large N, the total number of edges in the one-mode projected graph (the number of pairwise coactorships) is t times the number of pairs of actors in a new film, that is, $t\overline{n}(\overline{n}-1)/2$, while the total number of actors is $N = t\overline{m}$. Thus the mean degree of the one-mode projection network is

$$\langle k\rangle = \frac{\overline{n}(\overline{n}-1)}{\overline{m}}. \tag{9}$$

Therefore

$$\langle q\rangle = \frac{\langle k\rangle}{\overline{n}-1} = \frac{\overline{n}}{\overline{m}}. \tag{10}$$

As in the Simon model, the connections of this growing bipartite graph self-organize into a scale-free structure. Quite similarly to standard derivations for the Simon model, in the large network limit, the distribution takes the form [2,27]

$$Q(q) = (\gamma-1)B(q,\gamma), \tag{11}$$

where $B(,)$ is the β-function [42] and

$$\gamma = 2 + \frac{\overline{m}}{\overline{n}-\overline{m}}. \tag{12}$$

For $q \gg 1$, the asymptotics of $Q(q)$ is

$$Q(q) \sim (q+\gamma-1/2)^{-\gamma}, \tag{13}$$

so that γ is the exponent of the degree distribution. In the one-mode projection, this corresponds to

$$P(k) \sim [k+(\gamma-1/2)(\overline{n}-1)]^{-\gamma}. \tag{14}$$

That is, the projected degree distribution exhibits a power law behavior, with an offset $k_0=(\gamma-1/2)(\overline{n}-1)$. The presence of this offset, which may be large for large values of $\overline{n}$, can hinder the direct evaluation of the exponent γ. Therefore it is more appropriate to compare the degree distribution with the general expression Eq. (14).

To calculate the clustering coefficient, we need to recall our second assumption: If we consider a particular actor, i, who has played in q movies, in the thermodynamic limit none of his coactors repeats twice in different films. This means that in the projected network the triangles attached to a vertex i can only relate his coactors inside each separate movie. The number of such triangles is $q(\overline{n}-1)(\overline{n}-2)/2$, while the total number of possible triangles attached to i is $k(k-1)/2$ or, equivalently, $q(\overline{n}-1)[q(\overline{n}-1)-1]/2$. The local clustering as a function of the experience of an actor is then given by

$$c(q) = \frac{\overline{n}-2}{q(\overline{n}-1)-1}, \tag{15}$$

which, as a function of k, transforms into

$$c(k) = \frac{\overline{n}-2}{k-1}. \tag{16}$$

Then using the definition (6) readily yields the average clustering

$$\langle c\rangle = \sum_{k>1} P(k)c(k) = (\overline{n}-2)\sum_{k>1}\frac{P(k)}{k-1} = \frac{\overline{n}-2}{\overline{n}-1}\sum_{q>0}\frac{Q(q)}{q-1/(\overline{n}-1)}. \tag{17}$$

On the other hand, to compute the clustering coefficient c, defined in Eq. (4), we need to estimate the number of triangles attached to a vertex of degree k. As we have seen before, this number is $q(\overline{n}-1)(\overline{n}-2)/2$, or, in terms of k, $s(k)=k(\overline{n}-2)/2$. Therefore

$$\begin{aligned} c &= \frac{\sum_q P(q)q(n-1)(n-2)/2}{\sum_q P(q)q(n-1)[q(n-1)-1]/2} \\ &= \frac{(n-2)\langle q\rangle}{(n-1)\langle q^2\rangle - \langle q\rangle} \\ &= \frac{(n-2)\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}. \end{aligned} \tag{18}$$

One can see that the average clustering $\langle c\rangle$ converges to a finite value for any degree distribution, since the region of low degrees makes the main contribution. Consequently, Eq. (17) works well even if the degree distribution is fat-tailed. The clustering coefficient, on the other hand, approaches zero if the second moment of the degree distribution diverges. This divergence takes place for $\gamma \leqslant 3$ in the thermodynamic limit $(N,t\to\infty)$. In this case, c crucially depends on the degree cutoff k_c (or q_c) in the form $c \sim k_c^{-(3-\gamma)}$. Note that formula (18) may underestimate the value of the clustering coefficient whenever the degree distribution is fat-tailed and k_c is large.

V. RESULTS AND COMPARISON WITH REAL NETWORKS

To check the validity of our model, we proceed to compare the empirical data on collaboration networks with the predictions made in the previous section, as well as with numerical simulations of the model. The analytic predictions are specified in terms of two parameters, the average number of actors per collaboration act $\overline{n}$ and the average number of new actors $\overline{m}$. If all these actors are recruited at a constant rate, then we have in average $\overline{m}=N/t$ new actors per collaboration act. From these two parameters, using the results of the previous section, we can compute our predictions for all the properties of the networks described in Sec. III. When performing numerical simulations of the model, and in order to avoid discreteness, we use randomly distributed m and n. Their distributions are taken to be exponential with averages $\overline{m}$ and $\overline{n}$, respectively. This functional form corresponds to that of the distribution $S(n)$ empirically observed for actor and scientific collaboration networks [see Fig. 1(a)]. For the company directorship network on the other hand, we do not count with an empirical form of $S(n)$. Hence we checked both exponential and Poisson distributions. The global characteristics of the networks generated with this last distribution suit better their empirical counterparts. The results of the comparison between empirical data, theoretical predictions, and simulations are summarized in Table I.

We observe an agreement between the model predictions and the empirical results for the mean clustering $\langle c\rangle$. Note some deviations in the mean degree $\langle k\rangle$ of one-mode projected networks. These discrepancies are due to the fact that in our analysis we neglect the probability that some actors

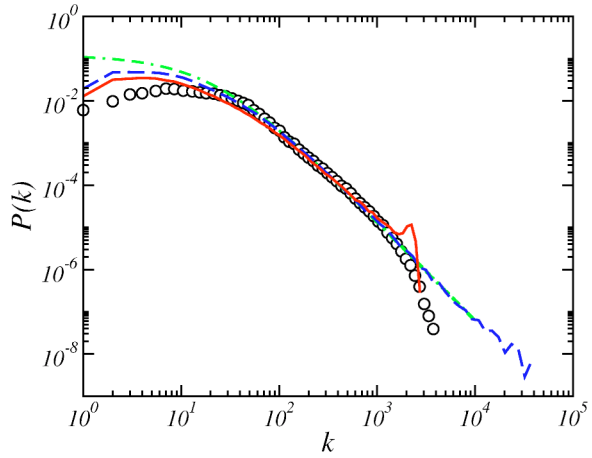

FIG. 5. Comparison among the degree distribution $P(k)$ for the empirical movie actor collaboration network (circles), for the theoretical prediction of Sec. IV (dot-dashed line), and for simulations of the original model (dashed curve) and the version with aging (solid line).

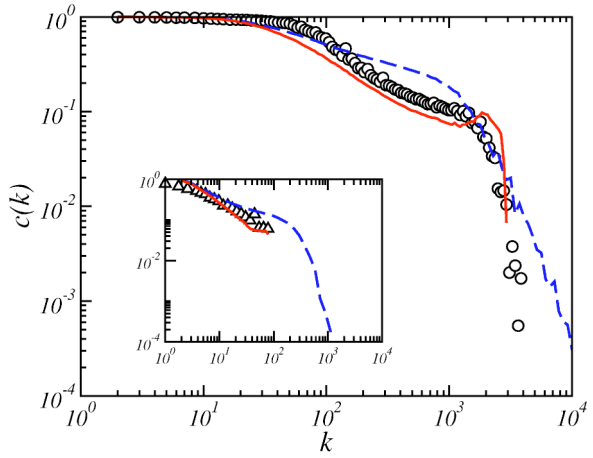

FIG. 6. Comparison among the clustering coefficient as a function of the degree $c(k)$ for the empirical movie actor collaboration network (circles), for the simulations of the original version of the model (dashed line), and for the model with aging (solid line). The main plot is for the actor costarring network and the inset for the scientific collaboration network.

for a new film have previously costarred in the same movies. For the net of codirectorships, the computed clustering coefficient c is in a reasonable agreement with the empirical value. In the other networks, however, the calculated values of c are severely underestimated. The reason for this is the poor quality of the approximate formula (18) for this model in the case of a fat-tailed degree distribution (see discussion in Sec. IV).

The exponents of the projected degree distribution are $\gamma=2.35$, $\gamma=2.43$, and $\gamma=4.7$ for the movie, coauthorship, and codirectorship networks, respectively. The exponent larger than 3 in this last case is compatible with the exponential decay observed empirically. The range of the empirical degree distribution in the coauthorship network is too small to compare with the asymptotic expression Eq. (14). So, we make this comparison only in the case of the movie actors network, Fig. 5. As can be seen, the agreement between the theoretical and empirical distributions is notorious. Only at very small and very large values of the degree a certain discrepancy can be noticed essentially due to the continuous degree approximation employed and the presence of a cutoff in the empirical distribution, respectively. This upper cutoff is inevitable due to two factors: (i) an actor physically cannot have an infinite number of costars, and (ii) finite size effects restrict the degrees of vertices (see, e.g., Ref. [41]).

The empirical local clustering and its analog obtained by numerical simulations (Fig. 6) demonstrate a more complex dependence on degree than the simple estimate of Eq. (16). From Fig. 6 we see that the function $c(k)$ computed from the model follows a slower decay than the corresponding empirical function for the movie actors network. The results for the mean degree of the nearest neighbors of a vertex as a function of its degree, $\bar{k}_{nn}(k)$, are represented in Fig. 7. Unexpectedly, apart from a small region for very small values of k, $\bar{k}_{nn}(k)$ decreases with degree, and the Pearson correlation coefficient is negative (see Table I). So that, unlike real-world collaboration graphs, networks with a fat-tailed degree distribution, generated by the simplest version of our model, show disassortative mixing. On the contrary, in the case of directorship networks, the model provides positive values for the Pearson coefficient r, in agreement with the empirical results, though a little lower (see discussion in the next section).

VI. SELF-ORGANIZED MODEL WITH AGING

At least in one aspect, the model presented above is a serious oversimplification of the mechanism underlying the growth of social collaboration networks. It allows an analytical treatment but leads to problems in the comparison with

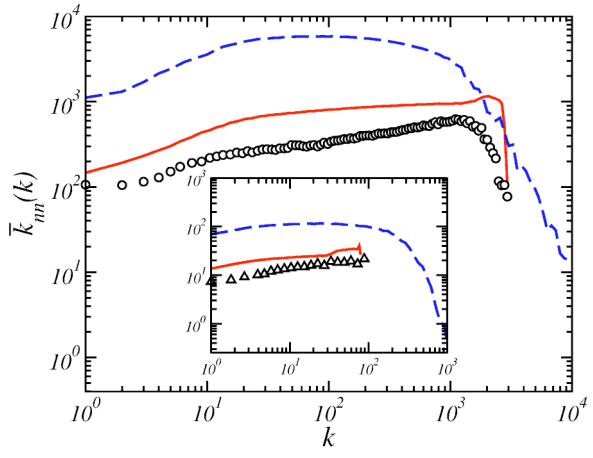

FIG. 7. Comparison among the average degree of the nearest neighbors as a function of the degree $\bar{k}_{nn}(k)$ for the empirical movie actor collaboration network (circles), the simulations of the original version of the model (dashed line), and for the model with aging (solid line). The main plot is for the actor costarring network and the inset for the scientific collaboration network.

the clustering coefficient and, especially, degree–degree correlations of empirical networks. The most important missing point is probably the aging of individual agents. This issue is evident in the case of the movie actor network (although it can be observed for the scientific collaboration network too). The Internet Movie Database site, from which the actor collaboration network was extracted, contains information spanning the whole century of the history of cinema, from Louis Lumiere to the most recent Hollywood productions. Considering that actors have a finite professional lifetime, it is unrealistic to allow them to take part in a movie irrespective of their age.

If we take into account this fact, two main consequences are immediately expected. On the one side, there should be an upper cutoff in the $Q(q)$ distribution corresponding to the professional life expectation of actors, as actually it is found in empirical distributions, and, in addition, not all actors may work together: only those who are contemporaneous. Obviously, this phenomenon affects much less the codirectorship network because of its exponential degree distribution.

To introduce this new ingredient in the model, we must first assume an aging rate for individual agents. The most straightforward way to do so is to suppose that the time is directly equivalent to the experience q. Actually, in more realistic situations, it may happen that each agent has its own aging rhythm. However, the latter version of aging would make the model more complex. Once time is identified, we must consider a survival probability distribution for agents. In parallel with biological systems, we will assume an almost sure survival until a certain age Q_0 and an exponential decay hereafter. The modification of the model then requires two new parameters: the cutoff Q_0 and the characteristic time of the exponential decay τ. The rest of the model remains the same. That is, in each step a new movie is produced, m actors are new and the rest of them $n-m$ are chosen at random with a probability proportional to their experience. In addition, we assume that the actors become inactive, i.e., they cannot be chosen again for new movies, with a probability given by the complementary of the survival distribution for their particular age q.

We carried out simulations with the new version of the model. Q_0 was fixed at 100 for the actors network and at 15 for scientific collaborations to agree with the cutoff observed in the empirical distributions $Q(q)$ of these networks. The value of the other parameter, τ, is not so easy to establish from phenomenological data, therefore we check several characteristic times. For the sake of concreteness, let us focus on the results obtained with $\tau=50$ for actor costarring and with $\tau=7$ for scientific coauthorships, which are realistic values compatible with the final decay of the $Q(q)$ empirical distributions. Actually, using τ two times bigger we did not observe essential differences in the properties of the networks. Moreover, a simple exponential survival probability (i.e., with the only parameter τ) also provides similar approximate values of the clustering coefficient and the Pearson coefficient. However, it does not allow one to satisfactory describe the whole degree distribution. Note that our choice of the aging parameters Q_0 and τ does not actually mean fitting of our final results, which are the clustering and degree–degree correlation characteristics. Indeed, the values of Q_0 and τ were chosen only to properly describe the degree distribution in the range of large degrees.

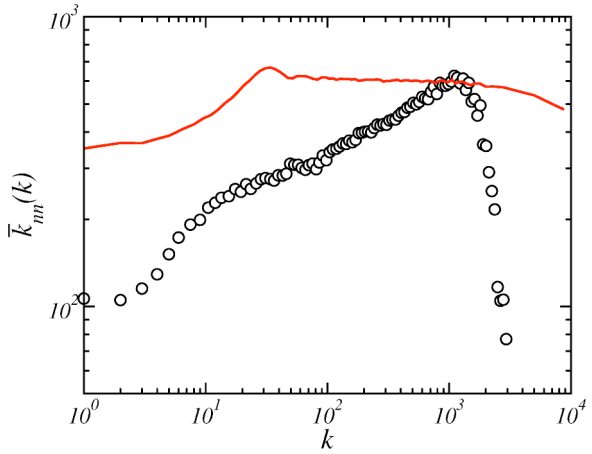

FIG. 8. Average degree of the nearest neighbors as a function of the degree $\bar{k}_{nn}(k)$ for an uncorrelated bipartite graph with the same degree distributions for both types of vertices as generated by our model (solid line). For comparison, the same quantity is displayed for the empirical actor network (circles).

In Fig. 6, the local clustering is plotted as a function of the degree for a network with aging and for the empirical actor network. The dependence $c(k)$ adjusts better to empirical data, and the computed clustering coefficients are closer to the empirical ones (see Table I). The improvement on these coefficients is understandable. As one can see from our simple analytical estimations, the direct introduction of the cutoff in the degree distribution seriously improves the values of the clustering coefficients.

A far more important point is that the aging changes the type of degree–degree correlations. In the version of the model with aging, the computed dependence of the mean degree of the nearest neighbor of a vertex on its degree properly describes the empirical dependence, as may be seen in Fig. 7. As a result, the computed values of the Pearson correlation coefficients turns out to be positive (assortative mixing) and close enough to the empirical values (see Table I). One should note that in the framework of the configuration model of an uncorrelated bipartite network, this agreement is impossible. We have checked this claim in the following way: We have measured the degree–degree correlations in the one-mode projection resulting from an uncorrelated bipartite graph with the same degree distributions for both types of vertices as generated by our model. In contrast to the self-organized model, see Fig. 8, the curve $\bar{k}_{nn}(k)$ turns out to be nearly flat, the Pearson coefficient being close to zero. This signals that the degree–degree correlations are practically absent in this case.

VII. CONCLUSION

In summary, we have studied a minimal model of evolving, self-organizing collaboration networks. This model is not based on a static perspective as was the configuration model, but on a dynamical mechanism to construct the net-

work. Besides, its basic constituents are preferential attachment and the bipartite structure of social networks. Our results show that the self-organized model offers a good starting point to explain existing empirical data. The model was compared with empirical results for a number of real networks, namely a network of scientific coauthorships, a network of movie actor collaborations, and a network of company codirectorships.

We have shown that, apart of a generic bipartite structure and the growth factor, one more element has to be taken into account in order to explain the empirical observations on the clustering and degree–degree correlations in collaboration networks. This key factor is the aging of collaborators. We demonstrate that in collaboration networks this effect is responsible for the positive (assortative) degree-degree correlations. We conclude that assortative mixing, which is generally observed in collaboration networks, is produced by the combination of their bipartite structure and the aging of the collaborators.

One should note that, in principle, even uncorrelated bipartite graphs (the configuration model) have correlated one-mode projections. However, the specific degree–degree correlations in these projections are quite weak. In other words, the configuration model graphs with degree distributions typical for movie actor nets show neither assortative nor disassortative mixing (they have $r \approx 0$). In contrast, our self-organized model provides correlated bipartite graphs, which, under natural assumptions, have one-mode projections with realistic structure and realistic correlations.

ACKNOWLEDGMENTS

We wish to thank M. E. J. Newman for making available the data for the cond-mat collaboration network, and M. Boguñá and A. Vespignani for helpful comments and discussions. Support from the Portuguese Research Council under Grant No. SFRH/BPD/14394/2003 and the European Commission Fet Open Project COSIN IST-2001-33555 is acknowledged. R.P.-S. acknowledges financial support from the Ministerio de Ciencia y Tecnología (Spain), and from the DURSI, Generalitat de Catalunya (Spain).

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **75**, 47 (2002).
[2] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[3] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, England, 2004).
[4] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[5] A. Wagner, Mol. Biol. Evol. **18**, 1283 (2001).
[6] H. Jeong, S. S. Mason, A. L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).
[7] R. V. Solé, R. Pastor-Satorras, E. Smith, and T. Kepler, Adv. Complex Syst. **5**, 43 (2002).
[8] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, ComPlexUs **1**, 38 (2003).
[9] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, England, 1994).
[10] G. Chartrand and L. Lesniak, *Graphs & Digraphs* (Wadsworth & Brooks/Cole, Menlo Park, 1986).
[11] R. Albert and A.-L. Barabási, Phys. Rev. Lett. **85**, 5234 (2000).
[12] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[13] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).
[14] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[15] A.-L. Barabási, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, and A. Schubert, Physica A **311**, 590 (2002).
[16] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[17] E. Ravasz *et al.*, Science **297**, 1551 (2002).
[18] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[19] A. Bekessy, P. Bekessy, and J. Komlos, Stud. Sci. Math. Hung. **7**, 343 (1972).
[20] E. A. Bender and E. R. Canfield, J. Comb. Theory, Ser. A **24**, 296 (1978).
[21] B. Bollobás, Eur. J. Comb. **1**, 311 (1980).
[22] N. C. Wormald, J. Comb. Theory, Ser. B **31**, 156 (1981); **31**, 158 (1981).
[23] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[24] J.-L. Guillaume and M. Latapy, cond-mat/0307095.
[25] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).
[26] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[27] H. A. Simon, Biometrika **42**, 425 (1955).
[28] If various kinds of collaborations are present, the collaboration network may be represented by a multipartite graph. Here we restrict ourselves to the simplest situation.
[29] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[30] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).
[31] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. E **65**, 066122 (2002).
[32] G. Szabó, M. Alava, and J. Kertesz, Phys. Rev. E **67**, 056102 (2003).
[33] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[34] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).
[35] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
[36] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, Proc. Natl. Acad. Sci. U.S.A. **99**, 2566 (2002).
[37] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, in *Statistical Mechanics of Complex Networks*, Vol. 625 of Lecture Notes in Physics, edited by R. Pastor-Satorras, J. M. Rubí, and A. Díaz-Guilera (Springer-Verlag, Berlin, 2003).

[38] Note, however, that $\overline{k}_{nn}(k)$ may be non-monotonous.
[39] Data available at http://www.nd.edu/~networks/database/index.html.
[40] Database located at http://xxx.lanl.gov/archive/cond-mat.
[41] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 205 (2004).
[42] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).