Evolution of the social network of scientific collaborations

A.L. Barabási[a,b,*], H. Jeong[a], Z. Néda[a,b,c], E. Ravasz[a], A. Schubert[d], T. Vicsek[b,e]

[a] *Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA*
[b] *Collegium Budapest, Institute of Advanced Study, Budapest, Hungary*
[c] *Department of Theoretical Physics, Babes-Bolyai University, Cluj, Romania*
[d] *Bibliometric Service, Library of the Hungarian Academy of Sciences, Hungary*
[e] *Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary*

Abstract

The co-authorship network of scientists represents a prototype of complex evolving networks. In addition, it offers one of the most extensive database to date on social networks. By mapping the electronic database containing all relevant journals in mathematics and neuro-science for an 8-year period (1991–98), we infer the dynamic and the structural mechanisms that govern the evolution and topology of this complex system. Three complementary approaches allow us to obtain a detailed characterization. First, empirical measurements allow us to uncover the topological measures that characterize the network at a given moment, as well as the time evolution of these quantities. The results indicate that the network is scale-free, and that the network evolution is governed by preferential attachment, affecting both internal and external links. However, in contrast with most model predictions the average degree increases in time, and the node separation decreases. Second, we propose a simple model that captures the network's time evolution. In some limits the model can be solved analytically, predicting a two-regime scaling in agreement with the measurements. Third, numerical simulations are used to uncover the behavior of quantities that could not be predicted analytically. The combined numerical and analytical results underline the important role internal links play in determining the observed scaling behavior and network topology. The results and methodologies developed in the context of the co-authorship network could be useful for a systematic study of other complex evolving networks as well, such as the world wide web, Internet, or other social networks. © 2002 Published by Elsevier Science B.V.

∗ Corresponding author. Tel.: +1-219-631-5767; fax: +1-219-631-5952.
  E-mail address: alb@nd.edu (A.L. Barabási).

1. Introduction

One of the most prolific mathematicians of all time, Paul Erdős has written over 1400 papers with over 500 co-authors. This unparalleled productivity inspired the concept of the Erdős number, which is defined to be one for his many co-authors, two for the co-authors of his co-authors and so on. The tightly interconnected nature of the scientific community is reflected by the conjecture that all publishing mathematicians, as well as many physicists and economists have rather small Erdős numbers [1]. Besides the immediate interest for scientometrics, the co-authorship networks is of general interest for understanding the topological and dynamical laws governing complex networks [2–17], as it represents the largest publicly available computerized social network.

Social networks have been much studied in social sciences [18,19]. A general feature of these studies is that they are restricted to rather small systems, and often view networks as static graphs, whose nodes are individuals and links represent various quantifiable social interactions.

In contrast, recent approaches with methodology rooted in statistical physics focus on large networks, searching for universalities both in the topology of the web and in the dynamics governing its evolution. These combined theoretical and empirical results have opened unsuspected directions for research and a wealth of applications in many fields ranging from computer science to biology [4,12–14,18,20–24]. Three important results seem to crystallize in this respect: First, most networks have the so-called *small world* property [2,19], which means that the average separation between the nodes is rather small, i.e., one can find a short path along the links between most pairs of nodes. Second, real networks display a degree of clustering higher than expected for random networks [2,4]. Finally, it has been found that the degree distribution contains important information about the nature of the network, for many large networks following a scale-free power-law distribution, inspiring the study of *scale-free* networks [3,5–8,12–14,24].

In addition to uncovering generic properties of real networks, these studies signal the emergence of a new set of modeling tools that considerably enhance our ability to characterize and model complex interactive systems. To illustrate the power of this these advances we choose to investigate in detail the collaboration network of scientists.

Recently, Newman has taken an important step towards applying modern network ideas to collaboration networks [10,11]. He studied several large database focusing on several fields of research over a 5-year period, establishing that collaboration networks have all the general ingredients of small world networks: they have a surprisingly short node-to-node distance and a large clustering coefficient [10], much larger than the one expected from a random Erdős–Rényi type network of similar size and average connectivity. Furthermore, the degree distribution appears to follow a power law [11].

Our study takes a different, but complementary approach to collaboration networks than that followed by Newman. We view collaboration networks as prototype of *evolving* networks, where the accent is on dynamics and evolution. Indeed, the co-authorship network constantly expands by the addition of new authors to the database, as well as

the addition of new internal links representing papers co-authored by authors that were already part of the database. The topological properties of these networks are determined by these dynamical and growth processes. Consequently, in order to understand their topology, we first need to understand the dynamical process that determines their evolution. In this aspect Newman's study focuses on the static properties of the collaboration graph, while our work investigates the dynamical properties of these networks. We show that such dynamical approach can explain many of the static topological features seen in the collaboration graph.

It is important to emphasize that the properties of the co-authorship network are not unique. The WWW is also a complex evolving network, where nodes and links are added (and removed) at a very high rate, the network topology being profoundly determined by these dynamical features [3,20,21,25]. The actor network of Hollywood is very similar to the co-authorship network, because it grows through the addition of new nodes (actors) and new links (movies linking existing actors) [2,4,14]. Similarly, the nontrivial scaling properties of many cellular [23], ecological [24] or business networks are all determined by dynamical processes that contributed to the emergence of these networks. So why single out the collaboration network as a case study? A number of factors have contributed to this choice. First, we needed a network for which the dynamical evolution is explicitly available. That is, in addition to a map of the network topology, it is important to know the time at which the nodes and links have been added to the network, crucial for revealing the network dynamics. This requirement reduces the currently available databases to two systems: the actor network, where we can follow the dynamics by recording the year of the movie release, and the collaboration network for which the paper publication year allows us to track the time evolution. Of these two, the co-authorship data is closer to a prototypical evolving network than the Hollywood actor database for the following reasons: in the science collaboration network the co-authorship decision is made entirely by the authors, i.e., decision making is delegated to the level of individual nodes. In contrast, for actors the decision often lies with the casting director, a level higher than the node. While in the long run this difference is not particularly important, the collaboration network is still closer in spirit to a prototypical evolving network such as social systems or the WWW.

Our work stands on three pillars. First, we use direct measurements on the available data to uncover the mechanism of network evolution. This implies determining the different parameters and uncovering the various competing processes present in the system. Second, building on the mechanisms and parameters revealed by the measurements we construct a model that allows us to investigate the large scale topology of the system, as well as its dynamical features. The predictions offered by a continuum theory of the model allow us to explain some of the results that were uncovered by ours, as well Newman's measurements. The third and final step will involve computer simulations of the model, serving several purposes: (i) It allows us to investigate quantities that could not be extracted from the continuum theory; (ii) Verifies the predictions of the continuum theory; (iii) Allows us to understand the nature of the measurements we can perform on the network, explaining some apparent discrepancies between the theoretical and the experimental results.
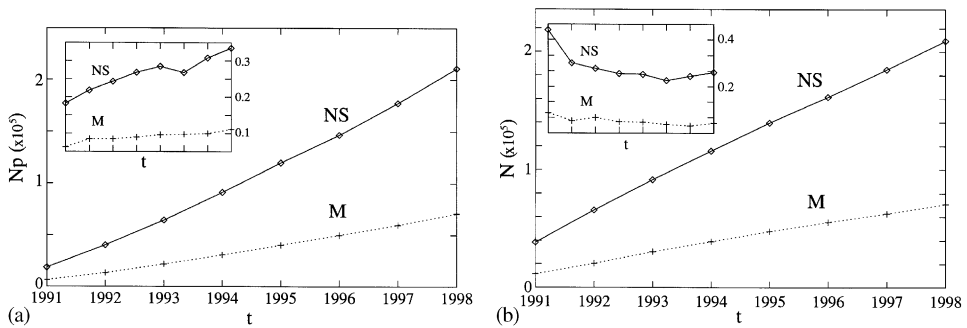
Fig. 1. (a) Cumulative number of papers for the M and NS databases in the period 1991–98. The inset shows the number of papers published each year. (b) Cumulative number of authors (nodes) for the M and NS databases in the period 1991–98. The inset shows the number of new authors added each year.

2. Databases: co-authorship in mathematics and neuro-science

For each research field whose practitioners collaborate in publications one can define a co-authorship network which is a reflection of the professional links between the scientists. In this network the nodes are the scientists and two scientists are linked if they wrote a paper together. In order to get information on the topology of a scientific co-authorship web one needs a complete dataset of the published papers, ideally from the birth of the discipline until today. However, computer databases cover at most the past several decades. Thus any study of this kind needs to be limited to only a recent segment of the database. This will impose unexpected challenges, that need to be addressed, since such limited data availability is a general feature of most networks.

The databases considered by us contain article titles and authors of all relevant journals in the field of mathematics (M) and neuro-science (NS), published in the period 1991–98. We have chosen these two fields for several reasons. A first factor was the size of the database: biological sciences or physics are orders of magnitude larger, too large to address their properties with reasonable computing resources. Second, the selected two fields offer sufficient diversity by displaying different publishing patterns: in NS collaboration is intense, while mathematics, although there is increasing tendency towards collaboration [26], is still a basically single investigator field.

In mathematics our database contains 70,975 different authors and 70,901 papers for an interval spanning 8 years. In NS the number of different authors is 209,293 and the number of published papers is 210,750. A complete statistics for the two considered database is summarized in Fig. 1, where we plot the cumulative number of papers and authors for the period 1991–98. We consider "new author" an author who was not present in the database from 1991 up to a given year.

Before proceeding we need to clarify a few methodological issues that affect the data analysis. First, in the database the authors are represented by their surname and initials of first and middle name, thus there is a source of error in distinguishing some of them. Two different authors with the same initials and surname will appear to be the same node in the database. This error is important mainly for scientists of Chinese and

Japanese descent. Second, seldom a given author uses one or two initials in different publications, and in such cases he/she will appear as separate nodes. Newman [10] showed that the error introduced by those problems is of the order of a few percents. Our results are also affected by these methodological limitations, but we do not expect that it will have a significant impact on our results.

3. Data analysis

In this section, we investigate the topology and dynamics of the two databases, M and NS. Our goal is to extract the parameters that are crucial to the understanding of the processes which determine the network topology, offering input for the construction of an appropriate model.

3.1. Degree distribution follows a power-law

A quantity that has been much studied lately for various networks is the degree distribution, $P(k)$, giving the probability that a randomly selected node has k links. Networks for which $P(k)$ has a power-law tail, are known as *scale-free* networks [3,13]. On the other hand, classical network models, including the Erdős–Rényi [27,28] and the Watts and Strogatz [4] models have an exponentially decaying $P(k)$ and are collectively known as *exponential* networks. The degree distributions of both the M and NS data indicate that collaboration networks are scale-free. The power-law tail is evident from the raw, uniformly binned data (Fig. 2a and b), but the scaling regime is better seen on the plot that uses logarithmic binning, reducing the noise in the tail (Fig. 2c). The cumulative data with logarithmic binning indicates $\gamma_M = 2.4$ and $\gamma_{NS} = 2.1$ for the two databases [29].[1]

We will see in the coming sections that the data indicates the existence of two scaling regimes with two different scaling exponents. The combination of these two regimes could easily give the impression of an exponential cutoff in the $P(k)$ for large k. Further analysis, offered in Sections 4–6, indicates that a consideration of two scaling regimes offers a more accurate description.

3.2. Average separation decreases in time

The ability of two nodes, i and j, to communicate with each other depends on the length of the shortest path, l_{ij}, between them. The average of l_{ij} over all pairs of nodes is denoted by $d = \langle l_{ij} \rangle$, and we call it the average separation of the network, characterizing the networks interconnectedness. Large networks can have surprisingly small separation, explaining the origin of the *small-world* concept [2,19]. Determining the average separation in a large network is a rather time-consuming procedure. Usually,

---

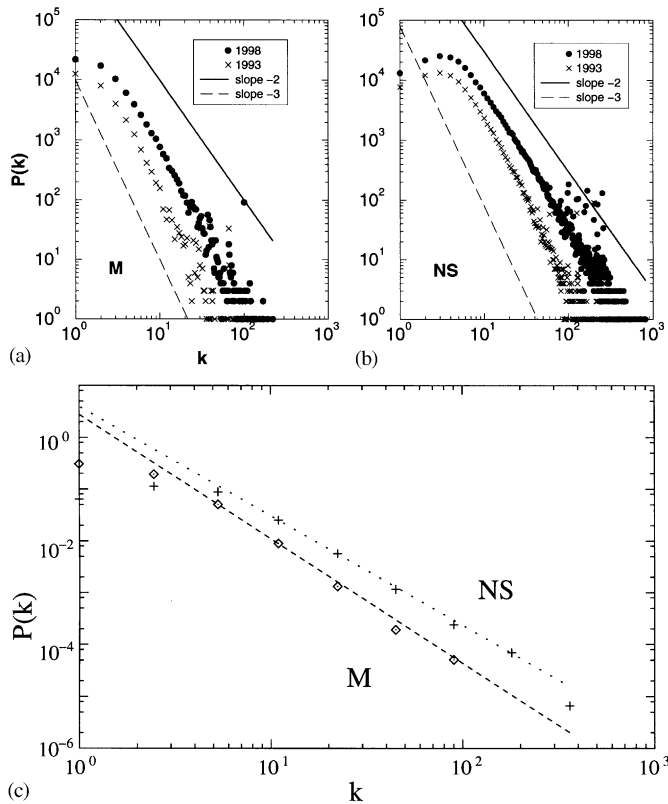[1] Note that such skewed distributions have been already modeled, from a different perspective, in scientometrics.

Fig. 2. Degree distribution for the (a) M and (b) NS database, showing the data based on the cumulative results up to years 1993 (×) and 1998 (●). (c) Degree distribution shown with logarithmic binning computed from the full dataset cumulative up to 1998. The lines correspond to the best fits, and have the slope 2.1 (NS, dotted) and 2.4 (M, dashed).

sampling a fraction of all nodes and determining their distance from all other points gives reasonable results. The results for the cumulative database are shown in Fig. 3.

The figure indicates that d decreases with time, which is highly surprising because all network models so far predict that the average separation should increase with system size [20,28]. The decreasing trend observed by us could have two different origins. First, it is possible that indeed, the separation does decrease as internal links, i.e., papers written by authors that were previously part of the database increase the interconnectivity thus decreasing the diameter. Second, the decreasing diameter could be a consequence of the fact that we have no access to the full database, but only starting from year 1991. As we demonstrate in Section 5, such incomplete dataset could result in an apparently decreasing separation even if otherwise for the full system the separation increases.

One can note the slow convergence of the diameter and the more connected nature of the NS field expressed by a smaller separation. The slow convergence indicates that
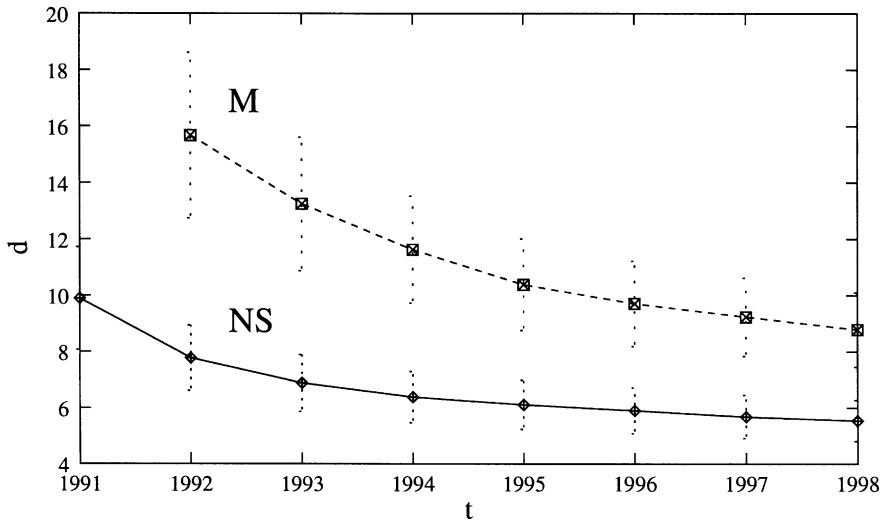
Fig. 3. Average separation in the M and NS databases. The separation is computed on the cumulative data up to the indicated year. The error bars indicate the standard deviation of the distances between all pairs of nodes.

perhaps even longer time interval is needed to reach the asymptotic limit, in which different relevant quantities take up a stationary value. The smaller separation for the NS field is expected, since mathematicians tend to work in smaller groups and write papers with fewer co-authors.

3.3. *Clustering coefficient decays with time*

An important phenomena characterizing the deviation of real networks from the completely random ER model is clustering. The clustering coefficient, a quantitative measure of this phenomena, C, can be defined as follows [4]: pick a node, i that has links to k_i other nodes in the system. If these k_i nodes form a fully connected clique, there are $k_i(k_i - 1)/2$ links between them, but in reality we find much fever. Let us denote by N_i the number of links that connect the selected k_i nodes to each other. The clustering coefficient for node i is then $C_i = 2N_i/k_i(k_i - 1)$. The clustering coefficient for the whole network is obtained by averaging C_i over all nodes in the system, i.e., $C = \langle C_i \rangle_i$. In simple terms the clustering coefficient of a node in the co-authorship network tells us how much a node's collaborators are willing to collaborate with each other, and it represents the probability that two of its collaborators wrote a paper together. The clustering coefficient for the cumulative network as a function of time is shown in Fig. 4.

The results, in agreement with the separation measurements, suggest a stronger interconnectedness for the NS compared with M, and a slow convergence in time to an asymptotic value.
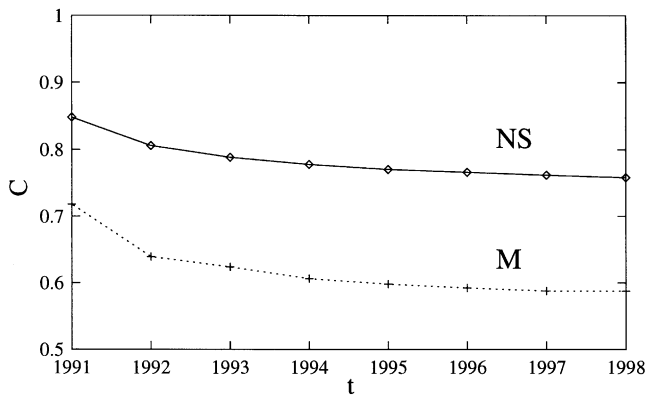
Fig. 4. Clustering coefficient of the M and NS database, determined for the cumulative data up to the year indicated on the *t*-axis.

3.4. Relative size of the largest cluster increases

It is important to realize that the collaboration network is fragmented in many clusters. There are several reasons for this. First, in every field there are scientists that do not collaborate at all, that is they are the only authors of all papers on which their name appears. This is more frequent in mathematics, which despite an increasing tendency toward collaboration [26], is still more fragmented than physics or neural science. Second, and most important, the database contains papers published only after 1990. Thus, there is a possibility that two authors co-authored a paper before 1990, but in our database they appear as disconnected.

If we look only at a single year, we see many isolated clusters of authors. The cumulative dataset containing several years develops a giant cluster, that contains a large fraction of the authors. To investigate the emergence of this giant connected component we measured the relative size of the largest cluster, *r*, giving the ratio between the number of nodes in the largest cluster and the total number of nodes in the system. A cluster is defined as a subset of nodes interconnected by links. Results from our cumulative co-authorship networks are presented in Fig. 5. As expected, in M the fraction of clustered researchers is considerably smaller than in NS.

The continuous increase in *r* may appear as the scenario commonly described as percolation [30] or the much studied emergence of the giant component in random networks [28]. However, the process leading to this giant cluster is fundamentally different from these much studied phenomena. In most research fields, apart from a very small fraction of authors that do not collaborate, all authors belong to a single giant cluster from the very early stages of the field. That is, the system is almost fully connected from the very first moment. The only reason why the giant cluster in our case grows so dramatically in the first several years is that we are missing the information on the network topology before 1991. A good example is the actor network, where the huge majority of the actors are part of the large cluster at any
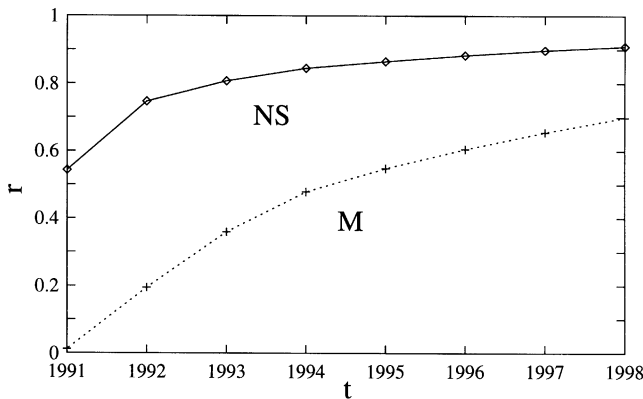
Fig. 5. Relative size of the largest cluster for the M and NS database. Results are computed on the cumulative data up to the given year.

stage of the network, starting from early 1900s until today. However, if we would start recording collaborations only after 1990 for example, the data would indicate, incorrectly, that many actors are disconnected. The increasing r indicates only the fact that we are reconstructing the already existing giant cluster, and it is only a partial measure of its emergence.

Finally, the fast convergence of the NS cluster size to an approximately stationary value around 0.9 indicates that after 1994 the network reached a roughly stationary topology, i.e., the basic alliances are uncovered. This does not seems to be the case for M, where after 10 years r still increases, perhaps due to smaller publication and collaboration rate in the field.

3.5. *Average degree increases*

With time the number of nodes in our co-authorship network increases due to arrival of new authors. The total number of links also increases through the connections made by new authors with old ones and by new connections between old authors. A quantity characterizing the network's interconnectedness is the average degree $\langle k \rangle$, giving the average number of links per author. The time dependence of $\langle k \rangle$ for the cumulative network is shown in Fig. 6, indicating an approximately linear increase of $\langle k \rangle$ with time. This is a rather important deviation from the majority of currently existing evolving network models, that assume a constant $\langle k \rangle$ as the network expands. As expected, the average degree for M is much smaller than for NS.

3.6. *Node selection is governed by preferential attachment*

Classical network models assume complete randomness, i.e., the nodes are connected to each other independent of the number of links they already had [2,27,28]. The discovery of the power-law connectivity distribution required the development of new
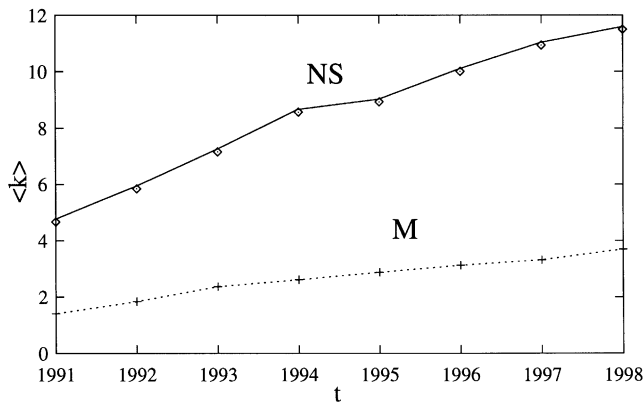
Fig. 6. Average number of links per node ($\langle k \rangle$) for the M and NS database. Results are computed on the cumulative data up to the given year.

modeling paradigms. A much used assumption is that in scale-free networks nodes link with higher probability to those nodes that already have a larger number of links, a phenomena labeled as preferential attachment [3,13]. Implicitly or explicitly, preferential attachment is part of all network models that aim to explain the emergence of the inhomogeneous network structure and power-law connectivity distribution [5–8]. The availability of dynamic data on the network development allows us to investigate its presence in the co-authorship network. For this network preferential attachment appears at two levels, that we discuss separately.

(i) *New nodes*: For a new author, that appears for the first time on a publication, preferential attachment has a simple meaning. It is more likely that the first paper will be co-authored with somebody that already has a large number of co-authors (links) that with somebody less connected. As a result "old" authors with more links will increase their number of co-authors at a higher rate than those with fever links. To investigate this process in quantitative terms we determined the probability that an old author with connectivity k is selected by a new author for co-authorship. This probability defines the $\Pi(k)$ distribution function. Calling "old authors" those present up to the last year, and "new author" those who were added during the last year, we determine the change in the number of links, Δk, for an old author that at the beginning of the last year had k links. Plotting Δk as a function of k gives the function $\Pi(k)$, describing the nature of preferential attachment. Since the measurements are limited to only a finite ($\Delta T = 1$ year) interval, we improve the statistics by plotting the integral of $\Pi(k)$:

$$\kappa(k) = \int_1^k \Pi(k') \, dk' \,. \tag{1}$$

If preferential attachment is absent, $\Pi(k)$ should be independent of k, as each node grows independently of its degree, and $\kappa(k)$ is expected to be linear. As Fig. 7 shows, we find that $\kappa(k)$ is nonlinear, increasing as $\kappa(k) \sim k^{v+1}$, where the best fit gives $v \simeq 0.8$ for M and $v \simeq 0.75$ for NS. This implies that $\Pi(k) \sim k^v$, where v is different from 1 [31]. As simulations have shown, such nonlinear dependence generates deviations
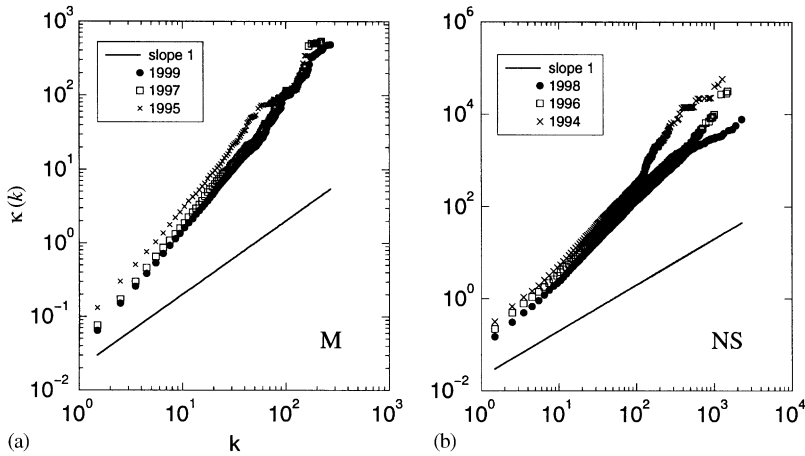
Fig. 7. Cumulated preferential attachment ($\kappa(k)$) of incoming new nodes for the M and NS database. Results computed by considering the new nodes coming in the specified year, and the network formed by nodes already present up to this year. In the absence of preferential attachment $\kappa(k) \sim k$, shown as continuous line on the figures.

from a power law $P(k)$ [31]. This was supported by analytical calculations [8], that demonstrated that the degree distribution follows a power law only for $v = 1$. The consequence of this nonlinearity will be discussed below.

(ii) *Internal links*: A large number of new links appear between old nodes as the network evolves, representing papers written by authors that were part of the network, but did not collaborate before. Such internal links are known to effect both the topology and dynamics of the network [5]. These internal links are also subject to preferential attachment. We studied the probability $\Pi(k_1, k_2)$ that an old author with k_1 links forms a new link with another old author with k_2 links. The $\Pi(k_1, k_2)$ probability map can be calculated by dividing $N(k_1, k_2)$, the number of new links between authors with k_1 and k_2 links, with the $D(k_1, k_2)$, number of pairs of nodes with connectivities k_1 and k_2 present in the system:

$$\Pi(k_1, k_2) = \frac{N(k_1, k_2)}{D(k_1, k_2)} \ . \tag{2}$$

The three-dimensional plot of $\Pi(k_1, k_2)$ is shown in Fig. 8, the overall behavior indicating preferential attachment: $\Pi(k_1, k_2)$ increases with as either k_1 or k_2's increase.

A natural hypothesis is to assume that $\Pi(k_1, k_2)$ factorizes into the product $k_1 k_2$. As Fig. 9 shows, we indeed find that

$$\kappa(k_1 k_2) = \int_1^{k_1 k_2} \Pi(k_1' k_2') \, d(k_1' k_2') \tag{3}$$

can be well approximated with a slope 2 as a function of $k_1 k_2$, indicating that for internal links the preferential attachment is linear in the degree.
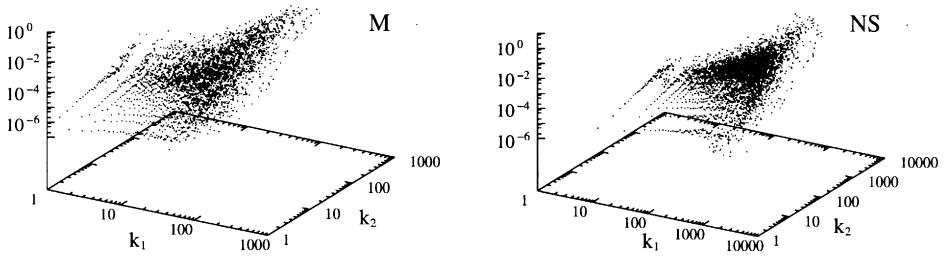
Fig. 8. Internal preferential attachment for the M and NS database, 3D plots: $\pi(k_1, k_2)$ as a function of k_1 and k_2. Results computed on the cumulative data in the last considered year.
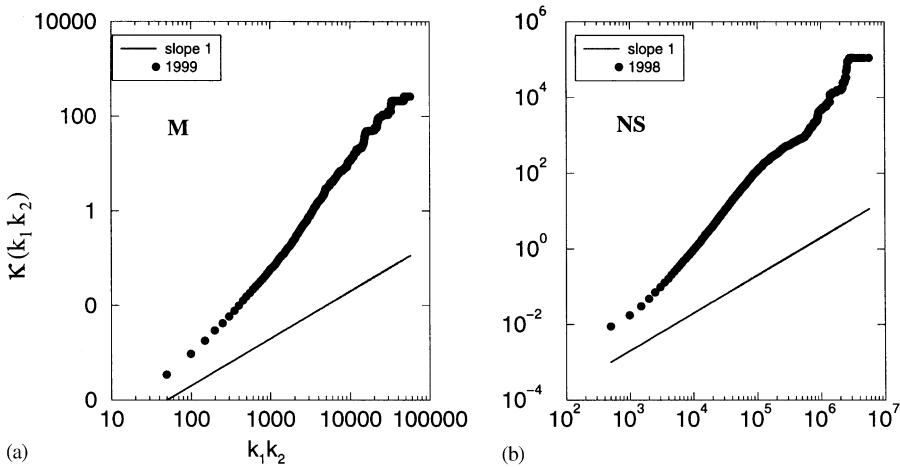

Fig. 9. Cumulated internal preferential attachment ($\kappa(k)$) for the M and NS database, scaling as a function of the $k_1 k_2$ product. Results computed on the cumulative data in the last considered year. The straight lines have slope 1, expected for $\kappa(k_1 k_2)$ if there would be no preferential attachment.

3.7. Modeling the web of science

In this section, we use the obtained numerical results to construct a simple model for the evolution of the co-authorship network. It is important to emphasize that the purpose of the model is to capture the main mechanisms that affect the evolution and the scaling of the network, and not to incorporate every numerical detail of the measured web. However, the advantage of the proposed model is its flexibility: features, neglected here, can be incorporated into the current modeling framework.

We denote by $k_i(t)$ the number of links node i has at time t; by $T(t)$ and $N(t)$ the total number of links and total number of nodes at time t, respectively.

We assume that all nodes present in the system are active, i.e., they can author further papers. This is a reasonable assumption as the time-span over which data is available to us is shorter than the professional lifetime of a scientist. In agreement with

Fig. 1, we consider that new researchers join the field at a constant rate, leading to

$$N(t) = \beta t \, . \tag{4}$$

The average number of links per node in the system at time t is thus given by

$$\langle k \rangle = \frac{T(t)}{N(t)} \, . \tag{5}$$

Fig. 9 suggests that the probability to create a new internal link between two existing nodes is proportional with the product of their connectivities. Consequently, denoting by a the number of newly created internal links per node in unit time, we write the probability that between node i and j a new internal link is created as

$$\Pi_{ij} = \frac{k_i k_j}{\sum'_{s,m} k_s k_m} \, 2N(t)a \, , \tag{6}$$

where the prime sign indicates that the summation is done for $s \neq m$ values.

The measurements also indicated (Fig. 7) that new nodes link to the existing nodes with preferential attachment, $\Pi(k)$ follows k^ν with $\nu \simeq 0.75 - 0.8$. Aiming to obtain an analytically solvable model, at this point we neglect this nonlinearity and we approximate $\Pi(k)$ with a linear k dependence. The effect of the nonlinearities will be discussed in Section 6. Thus, if node i has k_i links, the probability that an incoming node will connect to it is given by

$$\Pi_i = b \, \frac{k_i}{\sum_j k_j} \, , \tag{7}$$

where b is the average number of new links that an incoming node creates.

We have thus formulated the dynamical rules that govern our evolving network model, capturing the basic mechanism governing the evolution of the co-authorship network:

(1) Nodes join the network at a constant rate.
(2) Incoming nodes link to the already present nodes following preferential attachment (7).
(3) Nodes already present in the network form new internal links following preferential attachment (6).
(4) We neglect the aging of nodes, and assume that all nodes and links present in the system are active, able to initiate and receive new links.

In the model, we assume that the number of authors on a paper, m, is constant. In reality m is a stochastic variable, as the number of authors varies from paper to paper. However, for the scale-free model the exponent γ is known to be independent of m, thus making m a stochastic variable is not expected to change the scaling behavior.

4. Continuum theory

Taking into account that new links join the system with a constant rate, β, the continuum equation for the evolution of the number of links node i has can be

written as

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = \frac{b\beta k_i}{\sum_j k_j} + 2N(t)a\sum_j{}' \frac{k_i k_j}{\sum'_{s,m} k_s k_m} \ . \tag{8}$$

The first term on the right-hand side describes the contribution due to new nodes (7) and the second term gives the new links created with already existing nodes (6). The total number of links at time t can be computed taking into account the internal and external preferential attachment rules:

$$\sum_i k_i = T(t) = \int_0^t 2[N(t')a + b\beta]\,\mathrm{d}t' = t\beta(at + 2b) \ . \tag{9}$$

Consequently, the average number of links per node increases linearly in time,

$$\langle k \rangle = at + 2b \tag{10}$$

in agreement with our measurements on the collaboration network (Fig. 6). The master equation (8) can be solved if we approximate the double sum in the second term. Taking into account that we are interested in the asymptotic limit where the total number of nodes is large relative to the connectivity of the nodes, we can write

$$\sum_{s,m}{}' k_s k_m = \sum_s k_s \sum_m k_m - \sum_m k_m^2 \approx \left(\sum_i k_i\right)^2 \ . \tag{11}$$

We have used here the fact that $T(t)^2$ depends on N^2, while $\sum_i k_i^2$ depends only linearly on N (we investigate the $N \to \infty$ limit). Using (11) Eq. (8) now becomes

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = \frac{bk_i}{t(at+2b)} + \frac{2k_i a}{at+2b} \ . \tag{12}$$

Introducing the notation $\alpha = a/b$, we obtain

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = \frac{k_i}{t}\frac{2t\alpha + 1}{t\alpha + 2} \ . \tag{13}$$

This differential equation is separable, the general solution having the form

$$k_i(t) = C_i\sqrt{t}\sqrt{(2 + \alpha t)^3} \ . \tag{14}$$

The C_i integration constant can be determined from the initial conditions for node i. Since node i joins the system at time t_i, we have $k_i(t_i) = b$, leading to

$$k_i(t) = b\sqrt{\frac{t}{t_i}}\sqrt{\left(\frac{2 + \alpha t}{2 + \alpha t_i}\right)^3} \ . \tag{15}$$

This implies that for large times ($t \to \infty$) the connectivity of the node scales quadratically with time, i.e., $k(t) \sim t^2$.

A quantity of major interest is the degree distribution, $P(k)$. The nodes join the system randomly at a constant rate, which implies that the t_i values are uniformly distributed in time between 0 and t. The distribution function for the t_i in the $[0, t]$ interval is simply $\rho(t) = 1/t$. Using (15), $P(k)$ can be obtained after determining the $t_i(k_i)$ dependence from (15), giving

$$P(k) = -\rho(t) \left. \frac{\mathrm{d}t_i}{\mathrm{d}k_i} \right|_k \tag{16}$$

$$= \frac{1}{tk} \frac{S(k)(2 + \alpha S(k))}{2S(k) + 1} . \tag{17}$$

We denoted by $S(k)$ the function $t_i(k_i = k)$, determined from (15), solving the fourth-order polynomial equation

$$S(1 + \alpha S)^3 = \left(\frac{k_c}{k} \right)^2 \tag{18}$$

with

$$k_c = b\sqrt{t(2 + \alpha t)^3} . \tag{19}$$

In the $k \ll k_c$ limit $S \gg 1$ and we get

$$S(k) \approx \frac{1}{\alpha^3} \sqrt{\frac{k_c}{k}} . \tag{20}$$

For $k \gg k_c$ on the other hand $S \ll 1$, leading to

$$S(k) \approx \left(\frac{k_c}{k} \right)^2 . \tag{21}$$

Therefore, the scaling of $P(k)$ is controlled by a time-dependent crossover connectivity which is k_c. For $k \ll k_c$ the degree distribution will scale as

$$P(k) \propto k^{-3/2} \tag{22}$$

leading to the $\gamma = 3/2$ exponent. For $k \gg k_c$ we get instead

$$P(k) \propto k^{-3} \tag{23}$$

giving $\gamma = 3$. The crossover connectivity, k_c, increases linearly in time for $t \gg 2/\alpha$, which implies that in the asymptotic limit ($t \to \infty$) only the $\gamma = 3/2$ exponent is observable.

Note that this result predicts that the degree distribution has two scaling regimes, one with $\gamma = 3/2$ for small k, followed by a crossover to $\gamma = 3$ for large k. This crossover towards a larger exponent can be easily approximated with an exponential cutoff, which is why we believe that in Ref. [10] the power law with an exponential cutoff gave a reasonable fit. However, as Ref. [11] and our results show, for datasets with better statistics the two scaling regimes can be distinguished. Indeed, the crossover is visible in Fig. 2 as well, in particular for the degree distribution of NS. The degree distribution taken in 1993 has a clear $\gamma = 3$ tail, as for the studied short time-frame
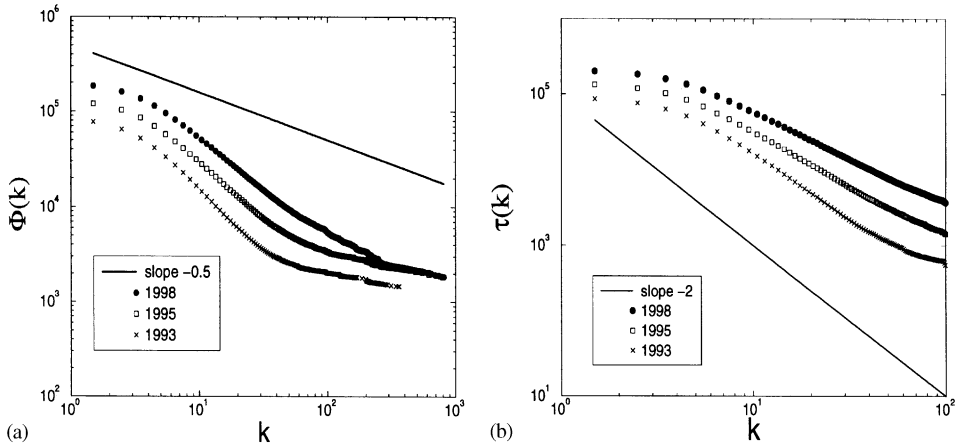
Fig. 10. Scaling of $\Phi(k)$ (a) and of $\tau(k)$ (b) for the NS database, demonstrating the trends in the small and large k behavior of the degree distribution (see text).

(3 years) k_c is expected to be low. This $\gamma = 3$ tail all but disappears, however, in 1998, being replaced with a $\gamma = 3/2$ exponent, as predicted by (22) for the limit $t \rightarrow \infty$. The M database shows similar characteristics, albeit the crossover is masked by a higher spread in the data point thanks to the weaker statistics.

Plotting instead of $P(k)$ two differently cumulated values, the $\gamma = 3/2$ and $\gamma = 3$ scaling regimes becomes more evident. Let us denote by $F(k)$ the primitive function of $P(k)$, defining

$$\Phi(k) = -F(1) - \int_1^k P(k')\,\mathrm{d}k' \,. \tag{24}$$

$\Phi(k)$ can be determined numerically by integrating $P(k)$ between 1 and k and subtracting the constant at which the integral converges. For small k the function $\Phi(k)$ should scale as

$$\Phi(k) \propto k^{-1/2} \tag{25}$$

assuming that $P(k)$ scales as given by (22). As Fig. 10a shows, we indeed find that for large t (1998) the measured $\Phi(k)$ function tends to approach the $k^{-1/2}$ behavior, which is less apparent on the small t curves (1993 and 1995).

To investigate the large k behavior of $P(k)$, we measured the $\tau(k)$ function defined as

$$\tau(k) = \int_k^\infty P(k')\,\mathrm{d}k' \,, \tag{26}$$

which captures the scaling of the tail. According to (23) for large k and small t one should observe

$$\tau(k) \propto k^{-2} \,. \tag{27}$$

As Fig. 10 shows, we indeed find that for NS at small t (1993) the large k scaling follows the prediction (26), and, as predicted, the scaling increasingly deviates from it as time increases.

5. Monte Carlo simulations

While the continuum theory discussed in the previous section predicts the connectivity distribution in agreement with the empirical data, there are other quantities, such as the node separation and clustering coefficient, that cannot be calculated using this method at this point. To investigate the behavior of these measures of the network topology next we study the model proposed in Section 3.6 using Monte Carlo simulations.

Due to memory and computing time limitations we investigated relatively small networks, with total number of nodes $N < 4000$. While these networks are considerably smaller than the real networks, their scaling and topological features should be representative. In order to form a reasonable number of internal links, we increased the parameter a in Eq. (6). For comparison purposes we note that in the real system we have $a_M = 0.31/(\text{year}) \simeq 10^{-4}/\text{simulation step}$ and $a_{NS} = 0.98/(\text{year}) \simeq 3.684 \times 10^{-5}/\text{simulation step}$, numbers that can be derived from the data shown in Figs. 6 and 1b.

The advantage of the modeling efforts, including the Monte Carlo simulations, is that they reproduce the network dynamics from the very first node. In contrast, the database we studied records nodes and links only after 1991, when much of the networks structure was already in place. By collecting data over several years we gradually discovered the underlying structure. We expect that after a quite long measurement time the structure revealed by the collected data will be statistically indistinguishable from the full collaboration network. However, the dynamics we measure during this process for the relevant quantities (diameter, average connectivity, clustering coefficient) might differ from those characterizing the full network, since all of them are computed on the *incomplete* network (revealed by the available data). However, Monte Carlo simulations allow us to investigate the effect of the data incompleteness on the relevant network measures.

We investigated the time dependence of the average connectivity, the diameter and the clustering coefficient, using the parameters $N_{max} = 1000$, $a = 0.001$, $\beta = 1$ and $b = 2$. In order to increase the statistics, the results were averaged over 10 independent configurations.

Average degree: As Fig. 11 indicates, asymptotically the average connectivity increases linearly, in agreement with both our measurements (see Fig. 6) and the continuum theory (see Eq. (5)).

Average separation: The empirical results indicated (see Fig. 3) that the average separation decreases with time for both databases. In contrast, our simulations show a monotonically increasing d, in apparent disagreement with the real system.

Note that an increasing diameter agrees with measurements done on other models, including scale-free and exponential networks, that all predict an approximately
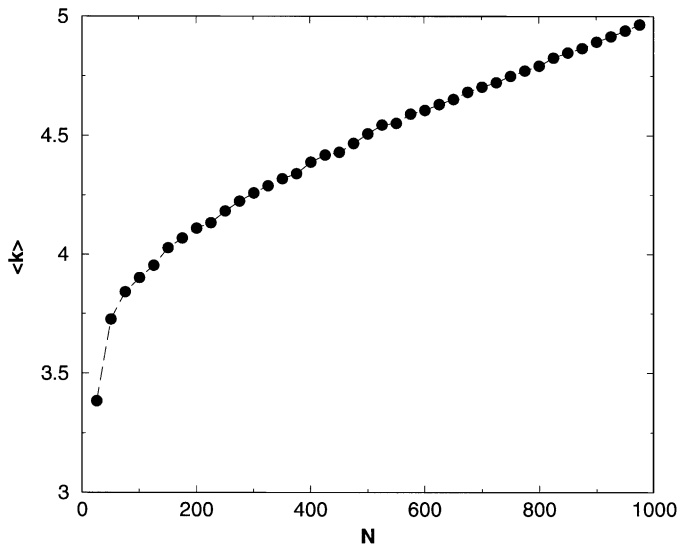
Fig. 11. Computer simulated dynamics of the average connectivity in the proposed model ($N_{max} = 1000$, $a = 0.001$, $\beta = 1$ and $b = 2$).

logarithmic increase with the number of nodes, $d \propto \ln(N)$ [28,32]. This contradiction between the models and our empirical data is rooted in the incomplete data we have for the first years of our measurements. To show this we perform the following simulation. We construct a network of $N = 1000$ nodes. However, we will record the apparent diameter of the network made of nodes that have been added only after a predefined time, mimicking the fact that the data available for us gives d only for publications after 1991. We find that the separation of this incomplete network has a decreasing tendency, slowly converging to the real value (Fig. 12), in agreement with the decrease observed in the empirical measurements (Fig. 3). This result underlies the importance of simulations in understanding the dynamics of complex networks, and resolves the conflict between the simulation and the empirical data. It also indicates that most likely the diameter of the M and NS database does increase in time, but such increase can be observed only if much longer time intervals will be available for study.

Clustering coefficient. The clustering coefficient predicted by our simulations is shown on Fig. 13. As the figure indicates, C depends strongly on the value of the parameter a. For $a = 0$ we have essentially the scale-free model [3] and the clustering coefficient has a monotonically decreasing tendency. For $a > 0$ however, the clustering coefficient decreases at the beginning and after reaching a minimum at N_{min} changes its course, asymptotically increasing with time. Thus, for all $a > 0$, we expect that in the asymptotic limit the clustering coefficient should increase, in agreement with our measurements on the collaboration network (see Fig. 4). The N_{min} position where the clustering coefficient has a minimum scales as power of the a parameter, as shown as the inset in Fig. 13.
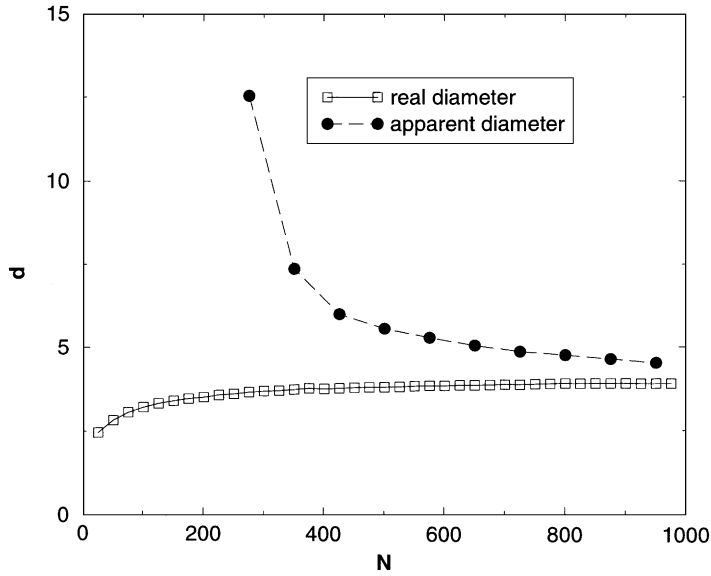
Fig. 12. Computer simulated dynamics for the real and apparently measured diameter value ($N_{max} = 1000$, $a = 0.001$, $\beta = 1$, $b = 2$ and $N_s = 200$).
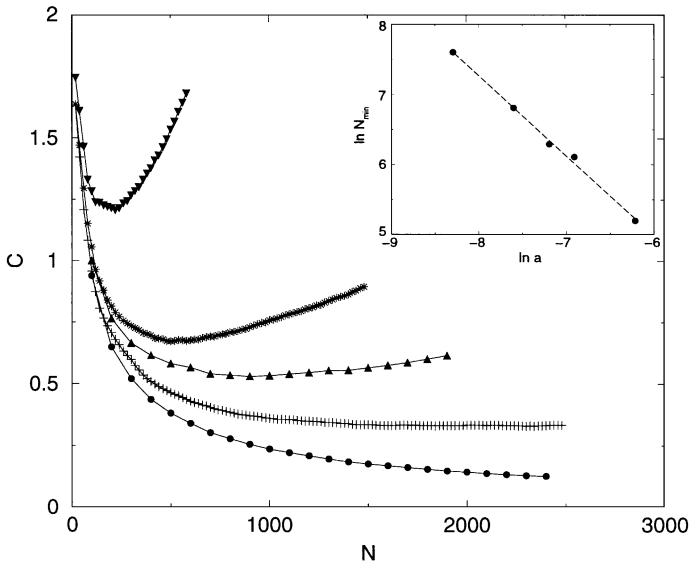

Fig. 13. Clustering coefficient for different values of the a parameter as a function of the system size N. ($N_{max} = 1000$, $\beta = 1$ and $b = 2$, values of a are 0 ($\bullet$), 0.00025 ($+$), 0.0005 ($\blacktriangle$), 0.00075 ($*$), 0.002 ($\blacktriangledown$).) The inset shows the scaling of the N_{min} value as a function of the a parameter. ($N_{max} = 1000$, $\beta = 1$ and $b = 2$, the line shows a fit $\ln N_{min} = -1.887 - 1.144 \ln a$.)
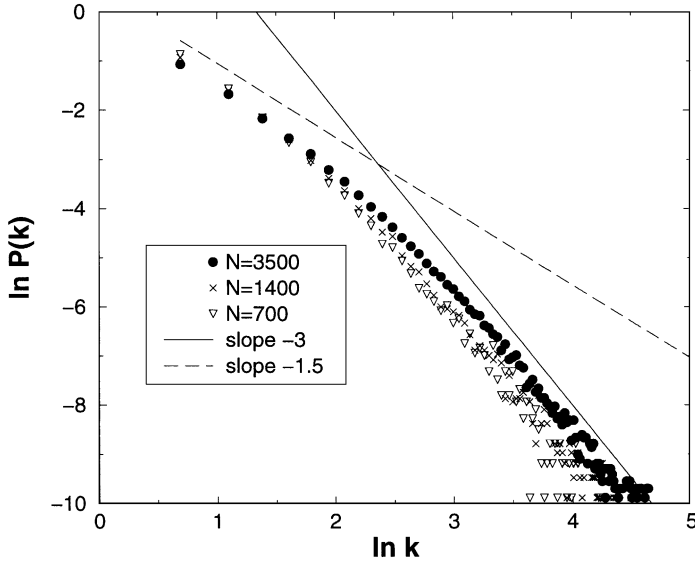
Fig. 14. Connectivity distributions as predicted by numerical simulation for different stages of evolution of the network ($a = 0.001$, $\beta = 1$ and $b = 2$).

We conclude thus that the decreasing C observed for our database, shown in Fig. 4, does not represent the asymptotic behavior. The observed behavior also indicates that one should view the values for C reported in the literature, and measured for finite time-frames (maximum 5 years) with caution, as they might not represent asymptotic values.

Degree distribution: The simulations provide $P(k)$ as well, allowing us to check the validity of the predictions of the continuum theory. Although the considered system sizes are rather small ($N_{\max} = 3500$) compared to the $N \to \infty$ approximation used in the analytical calculation and the $N_{\mathrm{M}} = 70,975$, $N_{\mathrm{NS}} = 209,750$ for the empirical data, the behavior of $P(k)$, shown in Fig. 14 agrees with our continuum model and measurements. For small k we observe the $\gamma = 1.5$ scaling, while for large k $P(k)$ converges to the predicted $\gamma = 3$ exponent.

6. Nonlinear effects

An issue that remained unresolved up to this point concerns the effect of the nonlinear preferential attachment. We have seen in Section 3.6 that for the incoming links we have

$$\Pi_i = b \, \frac{k_i^\nu}{\sum_j k_j^\nu} \tag{28}$$

with $\nu \approx 0.8$. On the other hand, for such preferential attachment Krapivsky et al. have shown that the degree distribution follows a stretched exponential, i.e., the power

law is absent [8]. This would indicate that $P(k)$ for the co-authorship network should follow a stretched exponential, which disagrees with our and Newman's findings (we have explicitly checked that a stretched exponential is not a good fit for our data). What could then override the known effect of the $v < 1$ nonlinear behavior? Next we propose a possible explanation: the linearity of the internal preferential attachment can restore the power-law nature of $P(k)$.

For non-integer v values the differential equation (8) governing the evolution of the connectivity is not analytically solvable. However, in the extreme case $v = 0$ (no preferential attachment for new nodes) the equation is again analytically tractable. Eq. (8) in this case has the form

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = \frac{b\beta}{N(t)} + 2N(t)a \sum_j \frac{k_i k_j}{\sum_{s,m} k_s k_m} \ . \tag{29}$$

Using $N(t) = \beta t$ and $\langle k \rangle = at + 2b$, which are valid in this case as well, following the steps described in Section 4, we obtain the differential equation:

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = \frac{b}{t} + \frac{2ak_i}{at + 2b} \ . \tag{30}$$

The general solution of this equation has the form:

$$k_i(t) = (2b + at)^2 C_i + \frac{2b + at}{4b}\left[2b + (2b + at)\ln\left(\frac{t}{2b + at}\right)\right] \ , \tag{31}$$

where C_i is an integration constant which can be determined using the $k_i(t_i) = b$ initial condition. The degree distribution cannot be determined analytically, since the $t_i(k_i)$ function is not analytical. However, taking the $\{t, t_i\} \to \infty$ limit, i.e., focusing on the network's long time evolution we obtain

$$k_i(t) \approx b\left(\frac{t}{t_i}\right)^2 \ , \tag{32}$$

which again predicts a power-law degree distribution

$$P(k) \propto k^{-3/2} \ . \tag{33}$$

Consequently, we obtain that in the asymptotic limit for $v = 0$ the scale-free degree distribution has the same tail as we obtained for $v = 1$. This result suggests that the linearity in the internal preferential attachment determines the asymptotic form of the degree distribution. The real exponent $v = 0.8$ is between the two asymptotically solvable cases $v = 0$ and 1, but, based on the limiting behavior of the two extremes we expect that independently of the value of $0 \leqslant v \leqslant 1$, in the asymptotic limit the degree distribution should converge to a power law with $\gamma = 3/2$. On the other hand, we expect that the nonlinear $v \neq 1$ behavior would have a considerable effect on the non-asymptotic behavior, which is not accessible analytically at this point.

To test further the potential effect of the nonlinearities, we have simulated the model discussed in Section 5 with $v = 0.75$, otherwise all parameters being unchanged. We show on Fig. 15 the degree distribution for the linear ($v = 1$) and the nonlinear ($v = 0.75$) case. As one can see, the $v = 1$ and 0.75 case can be hardly distinguished. This
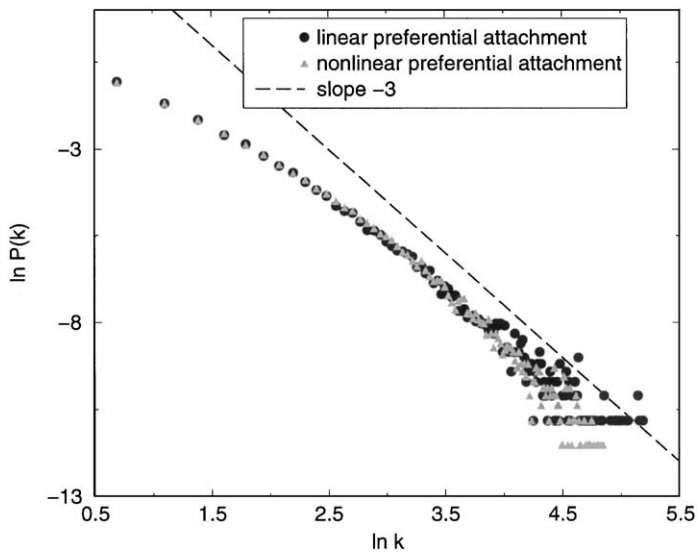
Fig. 15. Connectivity distribution generated by the numerical simulations for linear ($v = 1$) and nonlinear ($v = 0.75$) preferential attachment ($N_{\max} = 3500$, $a = 0.0005$, $\beta = 1$ and $b = 2$).

could have two origins. First, the simulations are limited to $t = 3500$ simulation steps, due to the discussed running time limitations. Thus we are hardly in the asymptotic regime. On the other hand, the agreement indicates that the nonlinear effect has no distinguishable effect on $P(k)$, again the internal attachment dominating the system behavior.

In summary, the domination of the internal attachment affects are expected to be even more dominant for the real network. Indeed, in the collaboration network the fraction of the links created as internal links is much higher than those created by the incoming nodes, as an author qualifies for a new incoming link only on his first paper. Most scientists contribute for a considerable time to the same filed, publishing numerous subsequent papers, and these later links will all appear as internal links. Thus, typically the number of internal links is much higher than the number of new links, the network's topology is much more driven by the internal links then by the external ones. This is one possible reason why the effect of the nonlinear behavior, while clearly present, cannot be detected in the functional form of $P(k)$.

7. Discussion

In the last 2 years we witnessed considerable advances in addressing the topology and the dynamics of complex networks. Along this road a number of quantities have been measured and calculated, aiming to characterize the network topology. However, most of these studies are fragmented, focusing on one or a few characteristics of the network at a time. Here, we presented a detailed study of a network of high

interest to the scientific community, the collaboration network of scientists, which also represents a prototype example of a complex evolving network. This study allows us to investigate to which degree can we use various known measures to characterize a given network. A first and important result of our investigation is that we need to be careful at distinguishing between the asymptotic and the intermediate behavior. In particular, most quantities used to characterize the network are time dependent. For example, the diameter, the clustering coefficient, as well as the average degree of the nodes are often used as basic time-independent network characteristics. Our empirical results show that many of these key quantities are time dependent, without a tendency to saturate within the available time-frame. Thus their value at a given moment tells us little about the network. They can be used, however, at any moment, to show that the network has small world properties, i.e., it has a small average separation, and a clustering coefficient that is larger than one expected for a random network.

A quantity that is often believed to offer a stationary measure of the network is the degree distribution. Our empirical data, together with the analytic solution of the model show that this is true only asymptotically for the co-authorship network; we uncover a crossover behavior between two different scaling regimes. We tend to believe that the model's predictions are not limited to the collaboration network; as on the WWW and for the actor collaboration network similar basic processes take place, chances are that similar crossovers could appear there as well.

A third important conclusion of the study regards the understanding that the measurements done on incomplete databases could offer trends that are opposite compared to that seen in the full system. An example is the node separation; we find that the empirically observed decreasing tendency is an artifact of the incomplete data. However, our simulations show that one can, with careful modeling, uncover such inconsistencies. But this also offers an important warning; for any network, before attempting to model it, we need to fully understand the limitations of the data collection process, and test their effect on the quantities of interest for us.

The model presented here represents only the starting point toward a complete modeling of collaborations in science. As we discussed throughout the paper, we have made several important approximations, sacrificing certain known network features for an analytical solution. For example, we neglected in our modeling effort the potential effect of aging [6,12], reflecting the fact that scientists retire and stop publishing papers. In the long run such aging effects will, undoubtedly, introduce exponential cutoffs in $P(k)$, as there are inherent limits on how many papers a researcher can write. Those effects, however, are not visible in our datasets. There are several potential reasons for this. Probably, the most important is that even the 8 years available to us for study is much shorter than the professional life of a scientist. Such aging induced cutoffs are expected to be visible only when time-frames of length of several time the scientist's professional life are studied. Data availability so far does not permit such studies.

A second simplification is that we assumed that each paper has exactly m authors. That is far from being so, as the numbers of co-authors varies greatly between papers. However, it is hard to imagine that the inclusion of a stochastic component in m would fundamentally affect our results. It is clear that such stochastic component will

not affect $P(k)$, and we feel that the effect on d or C is also negligible, but we lack at this point results to support this latter claim.

A surprising result of our study is the power-law character of $P(k)$, despite the fact that $\Pi(k)$ is nonlinear. We have shown that the existence of a linear internal attachment rule is able to restore the power law $P(k)$. Considering the fact that the largest fraction of links appear as internal links, compared with links created by new authors, it is fair to expect that the scaling determined by this internal linking process will dominate. The fact that for the two limits of the internal linking exponents, $v = 0$ and $v = 1$, we obtained power law $P(k)$ despite the nonlinear external $\Pi(k)$, suggests that such power law might appear for nonlinear, $v \neq 0, 1$ internal $\Pi(k)$ as well. Solving this problem is a formidable challenge, but it is perhaps worth the effort.

Finally, a more detailed modeling of the co-authorship network would involve the construction of bipartite graphs [33], in which we directly simulate the publishing of papers by several co-authors, which are all connected to each other. In such a model the basic unit is a paper, that involves several "old" and "new" authors. In such a framework one can simultaneously study the evolution of the co-authorship network (in which nodes are scientists linked by joint publications) and the publication network (in which nodes are papers, linked by joint authors). One can imagine that coupled continuum equations could be formulated for such bipartite network as well, which would eventually predict the network's dynamics and topology. Undoubtedly, including such detail in the modeling effort would increase the fidelity of the model. While challenging, following such path is beyond our goals here.

In summary, the modeling efforts presented here are only the starting point for a systematic investigation of the evolution of social networks. It is important to note that such modeling is open ended: more details can be incorporated, that could undoubtedly improve the agreement between the empirical data and theory. And such improvements might not be in vain; they could point towards a better understanding of the evolution of not only the co-authorship graph, but complex networks in general.

Acknowledgements

References

[1] http://www.oakland.edu/grossman/erdoshp.html.
[2] D.J. Watts, Small World, Princeton University Press, Princeton, NJ, 1999.
[3] A.L. Barabási, R. Albert, Science 286 (1998) 509.
[4] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 440.
[5] S.N. Dorogovtsev, J.F.F. Mendes, Europhys. Lett. 52 (2000) 33.

[6] S.N. Dorogovtsev, J.F.F. Mendes, Phys. Rev. E 62 (2000) 1842.
[7] S.N. Dorogovtsev, J.F.F. Mendes, Europhys. Lett. 50 (2000) 1; Phys. Rev. E 63 (2001) 025101;
    S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Phys. Rev. Lett. 85 (2000) 4633; cond-mat/0009090,
    2000; cond-mat/0011077, 2000.
[8] P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. 85 (2000) 4629;
    P.L. Krapivsky, S. Redner, Phys. Rev. E 63 (2001) 066123-1.
[9] S. Redner, Euro. Phys. J. B 4 (1998) 131.
[10] M.E.J. Newman, Proc. Natl. Acad. Sci. USA 98 (2001) 404.
[11] M.E.J. Newman, Phys. Rev. E 64 (2001) 016131.
[12] L.A.N. Amaral, A. Scala, M. Barthélémy, H.E. Stanley, Proc. Natl. Acad. Sci. USA 97 (2000) 11 149.
[13] A.L. Barabási, R. Albert, H. Jeong, Physica A 272 (1999) 173.
[14] R. Albert, A.L. Barabási, Phys. Rev. Lett. 85 (2000) 5234.
[15] R. Cohen, K. Erez, D. ben-Avraham, S. Havlin, Phys. Rev. Lett. 85 (2000) 4626.
[16] L. Kullmann, J. Kertész, cond-mat/0012410, 2000.
[17] R. Pastor-Satorras, A. Vespignani, Phys. Rev. Lett. 86 (2001) 3200.
[18] S. Wasserman, K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, 1994.
[19] M. Kochen (Ed.), The Small World, Ablex, Norwood, NJ, 1989.
[20] R. Albert, H. Jeong, A.L. Barabási, Nature 400 (1999) 130.
[21] S. Lawrence, C.L. Giles, Nature 400 (1999) 107.
[22] B.A. Huberman, L.A. Adamic, Nature 401 (1999) 131;
    M. Faloutsos, P. Faloutsos, C. Faloutsos, Proc. ACM SIGCOMM, Comput. Commun. Rev. 29 (1999)
    251;
    A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajalopagan, R. Stata, A. Tomkins, J. Wiener, in:
    Proceedings of the Ninth International World-Wide Web Conference, 2000.
[23] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.-L. Barabási, Nature 407 (2000) 651;
    D.A. Fell, A. Wagner, in: J.-H.S. Hofmeyr, J.M. Rohwer, J.L. Snoep (Eds.), Animating the Cellular
    Map, Stellenbosch University Press, Stellenbosch, 2000, p. 79;
    Technical Report 00-07-041, Santa Fe Institute, 2000.
[24] R.V. Solé, J.M. Montoya, cond-mat/0011196, 2000;
    J.M. Montoya, R.V. Solé, cond-mat/0011195, 2000.
[25] S. Lawrence, C.L. Giles, Science 280 (1998) 98.
[26] R. De Castro, J.W. Grossman, The Mathematical Intelligence 21 (3) (1999) 51–63.
[27] P. Erdös, A. Rényi, Publ. Math. Debrecen 6 (1959) 290.
[28] B. Bollobás, Random Graphs, Academic Press, London, 1985.
[29] A. Schubert, W. Glanzel, Scientometrics 6 (3) (1984) 149.
[30] A. Bunde, S. Havlin, Fractals and Disordered Systems, 2nd Edition, Springer, Berlin, Heidelberg,
    New York, 1996.
[31] A.-L. Barabási, R. Albert, H. Jeong, Physica A 281 (2000) 69.
[32] R. Albert, H. Jeong, A.L. Barabási, Nature 406 (2000) 378.
[33] M.E.J. Newman, S.H. Strogatz, D.J. Watts, Phys. Rev. E 64 (2001) 026118.