Scientific collaboration networks. I. Network construction and fundamental results

M. E. J. Newman
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501
and Center for Applied Mathematics, Cornell University, Rhodes Hall, Ithaca, New York 14853
(Received 5 December 2000; revised manuscript received 1 February 2001; published 28 June 2001)

Using computer databases of scientific papers in physics, biomedical research, and computer science, we have constructed networks of collaboration between scientists in each of these disciplines. In these networks two scientists are considered connected if they have coauthored one or more papers together. We study a variety of statistical properties of our networks, including numbers of papers written by authors, numbers of authors per paper, numbers of collaborators that scientists have, existence and size of a giant component of connected scientists, and degree of clustering in the networks. We also highlight some apparent differences in collaboration patterns between the subjects studied. In the following paper, we study a number of measures of centrality and connectedness in the same networks.

DOI: 10.1103/PhysRevE.64.016131 PACS number(s): 89.75.Hc, 89.65.−s, 89.70.+c, 01.30.−y

I. INTRODUCTION

A social network [1,2] is a set of people or groups each of which has connections of some kind to some or all of the others. In the language of social network analysis, the people or groups are called ''actors'' and the connections ''ties.'' Both actors and ties can be defined in different ways depending on the questions of interest. An actor might be a single person, a team, or a company. A tie might be a friendship between two people, a collaboration or common member between two teams, or a business relationship between companies.

Social network analysis has a history stretching back at least half a century, and has produced many results concerning social influence, social groupings, inequality, disease propagation, communication of information, and indeed almost every topic that has interested 20th century sociology. The *Physical Review* is, however, a physics journal. Why should a physicist be interested in social networks? There has, in fact, been a substantial surge of interest in social networks within the physics community recently, as evidenced by the large body of papers on the topic—see Refs. [3–24] and references therein. The techniques of statistical physics in particular turn out to be well suited to the study of these networks. Profitable use has been made of a variety of physical modeling techniques [5–7], exact solutions [8–13], Monte Carlo simulation [14–17], scaling and renormalization group methods [15–17], mean-field theory [18,19], percolation theory [20–22], the replica method [23], generating functions [20,22,24], and a host of other techniques familiar to the readers of this publication.

In this paper and the following one, we make use of some of these techniques in the study of some specific examples of social networks. However, our subject matter will be of interest to physicists for another reason: it's about them. In these two papers, we study networks in which the actors are scientists, and the ties between them are scientific collaborations, as documented by the learned articles that they write.

II. COLLABORATION NETWORKS

Traditional investigations of social networks have been carried out through field studies. Typically one looks at a fairly self-contained community such as a business community [25–27], a school [28,29], a religious or ethnic community [30], and so forth, and constructs the network of ties by interviewing participants, or by circulating questionnaires. A study will ask respondents to name those with whom they have the closest ties, probably ranked by subjective closeness, and may optionally call for additional information about those people or about the nature of the ties.

Studies of this kind have revealed much about the structure of communities, but they suffer from two substantial problems that make them poor sources of data for the kind of approach to network analysis that physics has adopted. First, the data they return are not numerous. Collecting and compiling data from these studies is an arduous process and most data sets contain no more than a few tens or hundreds of actors. It is a rare study that exceeds 1000 actors. This makes the statistical accuracy of many results poor, a particular difficulty for the large-system-size methods used in statistical physics. Second, they contain significant and uncontrolled errors as a result of the subjective nature of respondents' replies. What one respondent considers to be a friendship or acquaintance, for example, may be completely different from what another respondent does. In studies of schoolchildren [28,29], for instance, it is found that some children will claim friendship with every single one of their hundreds of schoolmates, while others will name only one or two friends. Clearly these respondents are employing different definitions of friendship.

Reliable statistics do exist for some other types of networks. Examples include the world-wide web [14,31,32], power grids [5], telephone call graphs [33], and airline timetables [34]. These graphs are certainly interesting in their own right, and furthermore might loosely be regarded as social networks, since their structure clearly reflects something about the structure of the society that created them. However, their connection to the ''true'' social networks discussed here is tenuous at best and so, for our purposes, they cannot offer a great deal of insight.

A more promising source of data is the affiliation network. An affiliation network is a network of actors connected by common membership in groups of some sort, such

as clubs, teams, or organizations. Examples that have been studied in the past include company CEOs and the clubs they frequent [26], company directors and the boards of directors on which they sit [25,35], women and the social events they attend [36], and movie actors and the movies in which they appear [5,34]. Data on affiliation networks tend to be more reliable than those on other social networks, since membership of a group can often be determined with a precision not available when considering friendship or other types of acquaintance. Very large networks can be assembled in this way as well, since in many cases group membership can be ascertained from membership lists, making time-consuming interviews or questionnaires unnecessary. A network of movie actors, for example, and the movies in which they appear has been compiled using the resources of the Internet Movie Database [37], and contains the names of nearly half a million actors—a much better sample on which to perform statistics than most social networks, although it is unclear whether this particular network has any real social interest.

In this paper we construct affiliation networks of scientists in which a link between two scientists is established by their coauthorship of one or more scientific papers. Thus the groups to which scientists belong in this network are the groups of coauthors of a single paper. This network is in some ways more truly a social network than many affiliation networks; it is probably fair to say that most pairs of people who have written a paper together are genuinely acquainted with one another, in a way that movie actors who appeared together in a movie may not be. There are exceptions—some very large collaborations, for example in high-energy physics, will contain coauthors who have never even met—and we will discuss these at the appropriate point. By and large, however, the network reflects genuine professional interaction between scientists, and may be the largest social network ever studied [38].

The idea of constructing a network of coauthorship is not new. Many readers will be familiar with the concept of the Erdős number, named after Paul Erdős, the Hungarian mathematician, one of the founding fathers of graph theory among other things [39]. At some point, it became a popular cocktail party pursuit for mathematicians to calculate how far removed they were in terms of publication from Erdős. Those who had published a paper with Erdős were given an Erdős number of 1, those who had published with one of those people but not with Erdős a number of 2, and so forth. The present author, for example, has an Erdős number of 3, via Robert Ziff and Mark Kac [40]. In the jargon of social networks, your Erdős number is the geodesic distance between you and Erdős in the coauthorship network. In recent studies [41–43], it has been found that the average Erdős number is about 4.7, and the maximum known finite Erdős number (within mathematics) is 15. These results are probably influenced to some extent by Erdős' prodigious mathematical output: he published at least 1512 papers, more than any other mathematician ever except possibly Leonhard Euler. However, quantitatively similar, if not quite so impressive, results are in most cases found if the network is centered on another mathematician. (On the other hand, the fifth most published mathematician, Lucien Godeaux, produced 644 papers, on 643 of which he was the sole author. He has no finite Erdős number [41]. Clearly sheer size of output is not a sufficient condition for high connectedness.)

There is also a substantial body of work in bibliometrics (a specialty within information science) on extraction of collaboration patterns from publication data—see Refs. [44–48] and references therein. However, these studies have not so far attempted to reconstruct entire collaboration networks from bibliographic data, concentrating more on organizational and institutional aspects of collaboration [49].

In this paper and the following one, we study networks of scientists using bibliographic data drawn from four publicly available databases of papers.

(1) Los Alamos e-Print Archive: a database of unrefereed preprints in physics, self-submitted by their authors, running from 1992 to the present. This database is subdivided into specialties within physics, such as condensed matter and high-energy physics.

(2) Medline: a database of articles on biomedical research published in refereed journals, stretching from 1961 to the present. Entries in the database are updated by the database's maintainers, rather than papers' authors, giving it relatively thorough coverage of its subject area. The inclusion of biomedicine is crucial in a study such as this one. In most countries biomedical research easily dwarfs civilian research on any other topic, in terms of both expenditure and human effort. Any study that omitted it would be leaving out the largest part of current scientific research.

(3) Stanford Public Information Retrieval System (SPIRES): a database of preprints and published papers in high-energy physics, both theoretical and experimental, from 1974 to the present. The contents of this database are professionally maintained. High-energy physics is an interesting case socially, having a tradition of much larger experimental collaborations than other disciplines.

(4) Networked Computer Science Technical Reference Library (NCSTRL): a database of preprints in computer science, submitted by participating institutions and stretching back about ten years.

We have constructed complete collaboration networks for each of these databases separately, and analyzed them using a variety of techniques, some standard, some invented for the purpose. A brief report of some of the work described here has appeared previously as Ref. [50].

III. FUNDAMENTAL RESULTS

For this study, we constructed collaboration networks using data from a five year period from 1995 to 1999 inclusive, although data for much longer periods were available in some of the databases. There were several reasons for using this fairly short time window. First, older data are less complete than newer for all databases. Second, we wanted to study the same time period for all databases, so as to be able to make valid comparisons between collaboration patterns in different fields. The coverage provided by both the Los Alamos Archive and the NCSTRL database is relatively poor before 1995, and this sets a limit on how far back we can look. Third, networks change over time, both because people

TABLE I. Some fundamental statistics for the scientific collaboration networks studied here. Numbers in parentheses are standard errors on the least significant figures.

		Los Alamos e-Print Archive					
	Medline	complete	astro-ph	cond-mat	hep-th	SPIRES	NCSTRL
Total number of papers	2163923	98502	22029	22016	19085	66652	13169
Total number of authors	1520251	52909	16706	16726	8361	56627	11994
First initial only	1090584	45685	14303	15451	7676	47445	10998
Mean papers per author	6.4(6)	5.1(2)	4.8(2)	3.65(7)	4.8(1)	11.6(5)	2.55(5)
Mean authors per paper	3.754(2)	2.530(7)	3.35(2)	2.66(1)	1.99(1)	8.96(18)	2.22(1)
Collaborators per author	18.1(1.3)	9.7(2)	15.1(3)	5.86(9)	3.87(5)	173(6)	3.59(5)
Size of giant component	1395693	44337	14845	13861	5835	49002	6396
First initial only	1019418	39709	12874	13324	5593	43089	6706
As a percentage	92.6(4)%	85.4(8)%	89.4(3)	84.6(8)%	71.4(8)%	88.7(1.1)%	57.2(1.9)%
2nd largest component	49	18	19	16	24	69	42
Clustering coefficient	0.066(7)	0.43(1)	0.414(6)	0.348(6)	0.327(2)	0.726(8)	0.496(6)

enter and leave the professions they represent and because practices of scientific collaboration and publishing change. In this particular study we have not examined time evolution in the network, although this is certainly an interesting topic for research and indeed is currently under investigation [51,52]. For our purposes, a short window of data is desirable, to ensure that the collaboration network is roughly static during the study.

The raw data for the networks described here are computer files containing lists of papers, including authors' names and possibly other information such as title, abstract, date, journal reference, and so forth. Construction of the collaboration networks is straightforward. The files are parsed to extract author names and as names are found a list is maintained of the ones seen so far—vertices already in the network—so that recurring names can be correctly assigned to extant vertices. Edges are added between each pair of authors on each paper. A naive implementation of this calculation, in which names are stored in a simple array, would take time $O(pn)$, where p is the total number of papers in the database and n the number of authors. This, however, turns out to be prohibitively slow for large networks since p and n are of similar size and may be a million or more. Instead therefore, we store the names of the authors in an ordered binary tree, which reduces the running time to $O(p \log n)$, making the calculation tractable, even for the largest databases studied here.

In Table I we give a summary of some of the basic results for the networks studied here. We discuss these results in detail in the rest of this section.

A. Number of authors

The size of the databases varies considerably from about a million authors for Medline to about ten thousand for NCSTRL. In fact, it is difficult to say with precision how many authors there are. One can say how many distinct *names* appear in a database, but the number of names is not the same as the number of authors. A single author may report their name differently on different papers. For example, F. L. Wright, Francis Wright, and Frank Lloyd Wright could all be the same person. Also, two authors may have the same name. Grossman and Ion [41] point out that there are two American mathematicians named Norman Lloyd Johnson, who are known to be distinct people but between whom computer programs such as ours cannot hope to distinguish. Even additional clues such as home institution or field of specialization cannot be used to distinguish such people, since many scientists have more than one institution or publish in more than one field. The present author, for example, has addresses at the Santa Fe Institute and Cornell University, and publishes in both statistical physics and paleontology.

In order to control for these biases, we constructed two different versions of each of the collaboration networks studied here, as follows. In the first, we identify each author by his or her surname and first initial only. This method is clearly prone to confusing two people for one, but will rarely fail to identify two names which genuinely refer to the same person. In the second version of each network, we identify authors by surname and all initials. This method can much more reliably distinguish authors from one another, but will also identify one person as two if they give their initials differently on different papers. Indeed this second measure appears to overestimate the number of authors in a database substantially. Networks constructed in these two different fashions therefore give upper and lower bounds on the number of authors, and hence also give bounds on many of the other quantities studied here. In Table I we give numbers of authors in each network using both methods, but for many of the other quantities we give only an error estimate based on the separation of the bounds.

B. Number of papers per author

The average number of papers per author in the various subject areas is in the range of around three to six over the five year period. The only exception is the SPIRES database, covering high-energy physics, in which the figure is significantly higher at 11.6. One possible explanation for this is that SPIRES is the only database that contains both preprints

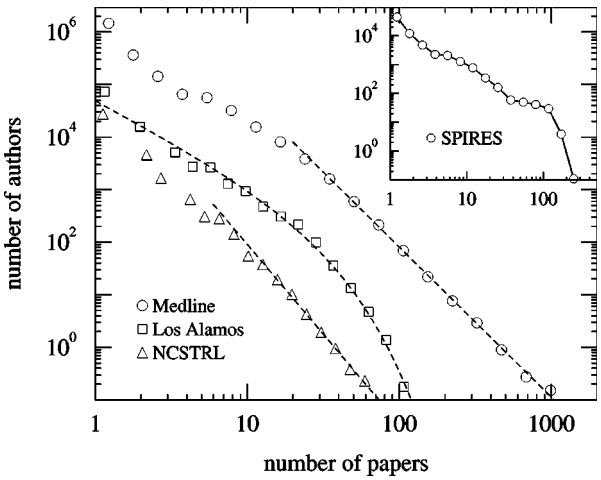

FIG. 1. Histograms of the number of papers written by authors in Medline, the Los Alamos Archive, and NCSTRL. The dotted lines are fits to the data as described in the text. Inset: the equivalent histogram for the SPIRES database.

and published papers. It is possible that the high figure for papers per author reflects duplication of papers in both preprint and published form. However, the maintainers of the database go to some lengths to avoid this [53], and a more probable explanation is perhaps that publication rates are higher for the large collaborations favored by high-energy physics, since a large group of scientists has more person-hours available for the writing of papers.

In addition to the average numbers of papers per author in each database, it is interesting to look at the distribution p_k of numbers k of papers per author. In 1926, Alfred Lotka [54] showed, using a data set compiled by hand, that this distribution followed a power law, with exponent approximately -2, a result that is sometimes referred to as Lotka's law of scientific productivity. In other words, in addition to the many authors who publish only a small number of papers, one expects to see a ''fat tail'' consisting of a small number of authors who publish a very large number of papers. In Fig. 1 we show on logarithmic scales histograms for each of our four databases of the numbers of papers published. (These histograms and all the others shown here were created using the ''all initials'' versions of the collaboration networks.) For the Medline and NCSTRL databases these histograms follow a power law quite closely, at least in their tails, with exponents of $-2.86(3)$ and $-3.41(7)$, respectively—somewhat steeper than those found by Lotka, but in reasonable agreement with other more recent studies [44,55,56]. For the Los Alamos Archive the pure power law is a poor fit. An exponentially truncated power law does much better:

$$p_k = Ck^{-\tau}e^{-k/\kappa}, \tag{1}$$

where τ and κ are constants and C is fixed by the requirement of normalization. (The probability p_0 of having zero papers is taken to be zero, since the names of scientists who have not written any papers do not appear in the database.) The exponential cutoff we attribute to the finite time window of five years used in this study, which prevents any one author from publishing a very large number of papers. Lotka and subsequent authors who have confirmed his law have not usually used such a window.

It is interesting to speculate why the cutoff appears only in physics and not in computer science or biomedicine. Surely the five year window limits everyone's ability to publish very large numbers of papers, regardless of their area of specialization? For the case of Medline one possible explanation is suggested by an inspection of the list of the most published authors: it transpires that most of these authors have names that are known to occur frequently. It is thus conceivable that these apparently highly published authors are really each several different people who have been conflated in our analysis, and hence that there is not after all any fat tail in the distribution, only the illusion of one produced by the large number of scientists with commonly occurring names. (This does not, however, explain why the tail appears to follow a power law.) This argument is strengthened by the sheer numbers of papers involved. For instance, the number 1 author in the Medline database published, it appears, 1697 papers, or about one paper a day, including weekends and holidays, every day for the entire course of our five year study. This seems to be an improbably large output.

Interestingly, the names that top the list in physics and computer science are not ones that are known to be common. Thus it is still unclear why the NCSTRL database should have a power-law tail, although this database is small and it is possible that it does possess a cutoff in the productivity distribution which is just not visible because of the limits of the data set.

For the SPIRES database, which is shown separately in the inset of the figure, neither pure nor truncated power law fits the data well, the histogram displaying a significant bump around the 100 paper mark. A possible explanation for this is that a small number of large collaborations published around this number of papers during the time period studied. Since each author in such a collaboration is then credited with publishing 100 papers, the statistics in the tail of the distribution can be substantially skewed by such practices.

C. Numbers of authors per paper

Grossman and Ion [41] have given results showing that the average number of authors on papers in mathematics has increased steadily over the last 60 years, from a little over 1 to its current value of about 1.5. Higher numbers still seem to apply to current studies in the sciences. Purely theoretical papers appear to be typically the work of two scientists, with high-energy theory and computer science showing averages of 1.99 and 2.22 authors per paper in our calculations. For databases covering experimental or partly experimental subject areas the averages are, not surprisingly, higher: 3.75 for biomedicine, 3.35 for astrophysics, 2.66 for condensed matter physics. The SPIRES high-energy physics database, however, shows the most startling results, with an average of 8.96 authors per paper, obviously a result of the presence of papers in the database written by very large collaborations. (Perhaps what is most surprising about this result is actually

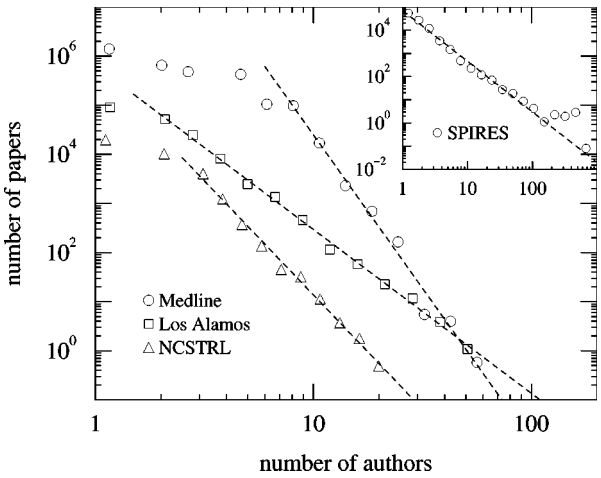

FIG. 2. Histograms of the number of authors on papers in Medline, the Los Alamos Archive, and NCSTRL. The dotted lines are the best fit power-law forms. Inset: the equivalent histogram for the SPIRES database, showing a clear peak in the 200 to 500 author range.

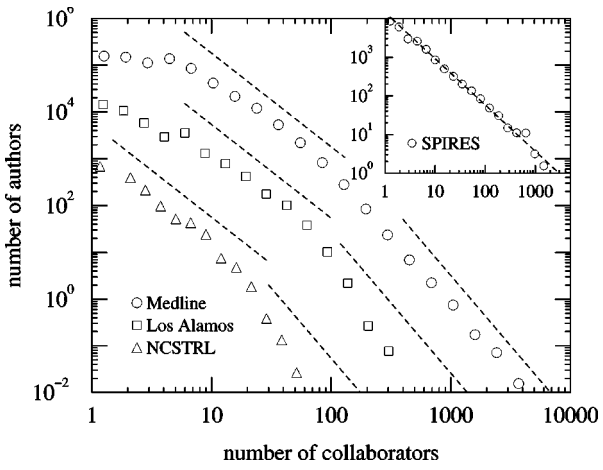

FIG. 3. Histograms of the number of collaborators of authors in Medline, the Los Alamos Archive, and NCSTRL. The dotted lines show how power-law distributions with exponents -2 and -3 would look on the same axes. Inset: the equivalent histogram for the SPIRES database, which is well fitted by a single power law (dotted line).

how small it is. The hundreds strong megacollaborations of CERN and Fermilab are sufficiently diluted by theoretical and smaller experimental groups that the number is only 9, and not 100.)

Distributions of numbers of authors per paper are shown in Fig. 2, and appear to have power-law tails with widely varying exponents of $-6.2(3)$ (Medline), $-3.34(5)$ (Los Alamos Archive), $-4.6(1)$ (NCSTRL), and $-2.18(7)$ (SPIRES). The SPIRES data, which are again shown in a separate inset, also display a pronounced peak in the distribution around 200–500 authors. This peak presumably corresponds to the large experimental collaborations that dominate the upper end of this histogram.

The largest number of authors on a single paper was 1681 (in high-energy physics, of course).

D. Numbers of collaborators per author

The differences between the various disciplines represented in the databases are emphasized still more by the numbers of collaborators that a scientist has, the total number of people with whom a scientist wrote papers during the five year period. The average number of collaborators is markedly lower in the purely theoretical disciplines (3.87 in high-energy theory, 3.59 in computer science) than in the wholly or partly experimental ones (18.1 in biomedicine, 15.1 in astrophysics). But the SPIRES high-energy physics database takes the prize once again, with scientists having an impressive 173 collaborators, on average, over a five year period. This clearly begs the question whether the high-energy coauthorship network can be considered an accurate representation of the high-energy physics community at all; it seems unlikely that many authors would know 173 colleagues well.

The distributions of numbers of collaborators are shown in Fig. 3. In all cases they appear to have long tails, but only the SPIRES data (inset) fit a power-law distribution well, with a low measured exponent of -1.20. Note also the small peak in the SPIRES data around 700—presumably again a result of the presence of large collaborations.

For the other three databases, the distributions show some curvature. This may, as we have previously suggested [50], be the signature of an exponential cutoff, produced once again by the finite time window of the study. Redner [57] has suggested an alternative origin for the cutoff using growth models of networks—see Ref. [10]. Another possibility has been put forward by Barabási [58], based on models of the collaboration process. In one such model [51], the distribution of the number of collaborators of an author follows a power law with slope -2 initially, changing to slope -3 in the tail, the position of the crossover depending on the length of time for which the collaboration network has been evolving. We show slopes -2 and -3 as dotted lines on the figure, and the agreement with the curvature seen in the data is moderately good, particularly for the Medline data. (For the Los Alamos and NCSTRL databases, the slope in the tail seems to be somewhat steeper than -3.)

E. Size of the giant component

In the theory of random graphs [24,59–61] it is known that there is a continuous phase transition with increasing density of edges in a graph at which a ''giant component'' forms, i.e., a connected subset of vertices whose size scales extensively. Well above this transition, in the region where the giant component exists, the giant component fills a large portion of the graph, and all other components (i.e., connected subsets of vertices) are small, with average size independent of the number n of vertices in the graph. We see a situation reminiscent of this in all of the graphs studied here: a single large component of connected vertices that fills the majority of the volume of the graph, and a number of much smaller components filling the rest. In Table I we show the size of the giant component for each of our databases, both as total number of vertices and as a fraction of system size.

In all cases the giant component fills around 80% or 90% of the total volume, except for high-energy theory and computer science, which give smaller figures. A possible explanation of these two anomalies may be that the corresponding databases give poorer coverage of their subjects. The hep-th high-energy database is quite widely used in the field, but overlaps to an extent with the longer established SPIRES database, and it is possible that some authors neglect it for this reason [53]. The NCSTRL computer science database differs from the others in this study in that the preprints it contains are submitted by participating institutions, of which there are about 160. Preprints from institutions not participating are mostly left out of the database, and its coverage of the subject area is, as a result, incomplete.

We also show in Table I the size of the second largest component in each of our networks. This component is in all cases far smaller than the giant component—typically consisting of only 20 or 30 authors—in qualitative agreement with our expectations from the theory of random graphs.

The figure of 80–90 % for the size of the giant component is a promising one. It indicates that the vast majority of scientists are connected via collaboration, and hence via personal contact, with the rest of their field. Furthermore, as we show in the following paper [62], the path through the network that connects two scientists is typically very short. Despite the prevalence of journal publishing and conferences in the sciences, person-to-person contact is still of paramount importance in the communication of scientific information, and it is reasonable to suppose that the scientific enterprise would be significantly hindered if scientists were not so well connected to one another.

F. Clustering coefficients

An interesting idea circulating in the social networks community currently is that of ''transitivity,'' which, along with its sibling ''structural balance,'' describes symmetry of interaction among trios of actors. ''Transitivity'' has a different meaning in sociology from its meaning in mathematics and physics, although the two are related. It refers to the extent to which the existence of ties between actors A and B and between actors B and C implies a tie between A and C. The transitivity, or more precisely the fraction of transitive triples, is that fraction of connected triples of vertices which also form ''triangles'' of interaction. Here a connected triple means an actor who is connected to two others. In the physics literature, this quantity is usually called the clustering coefficient C [5], and can be written

$$C = \frac{3 \times \text{number of triangles on the graph}}{\text{number of connected triples of vertices}}. \qquad (2)$$

The factor of 3 in the numerator compensates for the fact that each complete triangle of three vertices contributes three connected triples, one centered on each of the three vertices, and ensures that $C=1$ on a completely connected graph. On all random graphs $C=O(n^{-1})$ [5,24], where n is the number of vertices, and hence goes to zero in the limit of large graph size. In social networks it is believed that the clustering coefficient will take a nonzero value even in very large networks, because there is a finite (and probably quite large) probability that two people will be acquainted if they have another acquaintance in common. This is a hypothesis we can test with our collaboration networks. In Table I we show values of the clustering coefficient C, calculated from Eq. (2), for each of the databases studied, and as we see the values are indeed large, as large as 0.7 in the case of the SPIRES database, and around 0.3 or 0.4 for most of the others.

There are a number of possible explanations for these high values of C. First of all, it may be that they indicate simply that collaborations of three or more people are common in science. Every paper that has three authors clearly contributes a triangle to the numerator of Eq. (2) and hence increases the clustering coefficient. This is, in a sense, a trivial form of clustering, although it is by no means socially uninteresting.

In fact it turns out that this effect can account for some but not all of the clustering seen in our graphs. One can construct a random graph model of a collaboration network that mimics the trivial clustering effect, and the results indicate that only about a half of the clustering that we see is a result of authors collaborating in groups of three or more [24]. The rest of the clustering must have a social explanation, and there are some obvious possibilities.

(1) A scientist may collaborate with two colleagues individually, who may then become acquainted with one another through their common collaborator, and so end up collaborating themselves. This is the usual explanation for transitivity in acquaintance networks [1].

(2) Three scientists may all revolve in the same circles—read the same journals, attend the same conferences—and, as a result, independently start up separate collaborations in pairs, and so contribute to the value of C, although only the workings of the community, and not any specific personal interaction, is responsible for introducing them.

(3) As a special case of the previous possibility—and perhaps the most likely case—three scientists may all work at the same institution, and as a result may collaborate with one another in pairs.

Interesting studies could no doubt be made of these processes by combining our network data with data on, for instance, institutional affiliations of scientists. Such studies are, however, perhaps better left to the social scientists.

The clustering coefficient of the Medline database is worthy of brief mention, since its value is far smaller than those for the other databases. One possible explanation of this comes from the unusual social structure of biomedical research, which, unlike the other sciences, has traditionally been organized into laboratories, each with a ''principal investigator'' supervising a large number of postdoctoral associates, students, and technicians working on different projects. This organization produces a treelike hierarchy of collaborative ties. A tree has no loops in it, and hence no triangles to contribute to the clustering coefficient. Although the biomedicine hierarchy is certainly not a perfect tree, it may be sufficiently treelike for the difference to show up in the value of C. Another possible explanation comes from the

generous tradition of authorship in the biomedical sciences. It is common, for example, for a researcher to be made a coauthor of a paper in return for synthesizing reagents used in an experimental procedure. Such a researcher will in many cases have a less than average likelihood of developing new collaborations with their collaborators' friends, and therefore of increasing the clustering coefficient.

IV. CONCLUSIONS

In this paper we have studied social networks of scientists in which the actors are authors of scientific papers, and a tie between two actors represents coauthorship of one or more papers. Drawing on the lists of authors in four databases of papers in physics, biomedical research, and computer science, we have constructed explicit networks for papers appearing between the beginning of 1995 and the end of 1999. We have calculated a large number of statistics for our networks, including typical numbers of papers per author, authors per paper, and numbers of collaborators per author in the various fields. We note that the distributions of these quantities roughly follow a power-law form, although there are some deviations which may be due to the finite time window used for the study. We also note that in all the networks studied there exists a giant component of scientists any two of whom can be connected by a short path of intermediate collaborators.

A number of differences are apparent between the fields studied. Researchers in experimental disciplines are found to have larger numbers of collaborators on average than those in theoretical disciplines, with high-energy physicists having easily the largest average number of collaborators. We also find that in biomedicine the degree of network clustering is much lower than in other fields, possibly indicating differences in social organization between biomedical and other research communities.

In the following paper [62], we continue the study of the networks introduced here, looking at a variety of nonlocal network properties. Among other things, we look at the typical distances between pairs of scientists through the network, evaluate a number of centrality indices for our networks, and propose a method for calculating the strength of collaboration between scientists.

ACKNOWLEDGMENTS

The author would particularly like to thank Paul Ginsparg for his invaluable help in obtaining the data used for this study. The data were generously made available by Oleg Khovayko, David Lipman, and Grigoriy Starchenko (Medline), Paul Ginsparg and Geoffrey West (Los Alamos e-Print Archive), Heath O'Connell (SPIRES), and Carl Lagoze (NCSTRL). The Los Alamos e-Print Archive is funded by the NSF under Grant No. PHY-9413208. NCSTRL is funded through the DARPA/CNRI test suites program under DARPA Grant No. N66001-98-1-8908. The author would also like to thank László Barabási and Erzsébet Ravasz for making available an early version of Ref. [51], and László Barabási, Sankar Das Sarma, Paul Ginsparg, Rick Grannis, Jon Kleinberg, Laura Landweber, Sid Redner, Ronald Rousseau, Steve Strogatz, Duncan Watts, and Doug White for many useful comments and suggestions. This work was funded in part by the National Science Foundation and Intel Corporation.

[1] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[2] J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. (Sage Publications, London, 2000).
[3] M.E.J. Newman, J. Stat. Phys. **101**, 819 (2000).
[4] S.H. Strogatz, Nature (London) **410**, 268 (2001).
[5] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).
[6] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[7] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, in *Proceedings of the IEEE Symposium on Foundations of Computer Science* (IEEE, New York, 2000).
[8] C.F. Moukarzel, Phys. Rev. E **60**, 6263 (1999).
[9] S.N. Dorogovtsev and J.F.F. Mendes, Europhys. Lett. **50**, 1 (2000).
[10] P.L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).
[11] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).
[12] R.V. Kulkarni, E. Almaas, and D. Stroud, Phys. Rev. E **61**, 4268 (2000).
[13] J.M. Kleinberg, Nature (London) **406**, 845 (2000).
[14] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[15] M. Barthélémy and L.A.N. Amaral, Phys. Rev. Lett. **82**, 3180 (1999).
[16] M.E.J. Newman and D.J. Watts, Phys. Lett. A **263**, 341 (1999); Phys. Rev. E **60**, 7332 (1999).
[17] M.A. de Menezes, C.F. Moukarzel, and T.J.P. Penna, Europhys. Lett. **50**, 574 (2000).
[18] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[19] M.E.J. Newman, C. Moore, and D.J. Watts, Phys. Rev. Lett. **84**, 3201 (2000).
[20] C. Moore and M.E.J. Newman, Phys. Rev. E **61**, 5678 (2000); **62**, 7059 (2000).
[21] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).
[22] D.S. Callaway, M.E.J. Newman, S.H. Strogatz, and D.J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).
[23] A. Barrat and M. Weigt, Eur. Phys. J. B **13**, 547 (2000).
[24] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, e-print cond-mat/0007235.
[25] P. Mariolis, Soc. Sci. Q. **56**, 425 (1975).
[26] J. Galaskiewicz and P.V. Marsden, Soc. Sci. Res. **7**, 89 (1978).

[27] J.F. Padgett and C.K. Ansell, Am. J. Sociol. **98**, 1259 (1993).
[28] C.C. Foster, A. Rapoport, and C.J. Orwant, Behav. Sci. **8**, 56 (1963).
[29] T.J. Fararo and M. Sunshine, *A Study of a Biased Friendship Network* (Syracuse University Press, Syracuse, NY, 1964).
[30] H.R. Bernard, P.D. Kilworth, M.J. Evans, C. McCarty, and G.A. Selley, Ethnology **2**, 155 (1988).
[31] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Comput. Networks **33**, 309 (2000).
[32] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **406**, 378 (2000).
[33] J. Abello, A. Buchsbaum, and J. Westbrook, in *Proceedings of the Sixth European Symposium on Algorithms*, edited by G. Bilardi (Springer, Berlin, 1998).
[34] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11 149 (2000).
[35] G.F. Davis and H.R. Greve, Am. J. Sociol. **103**, 1 (1997).
[36] A. Davis, B.B. Gardner, and M.R. Gardner, *Deep South* (University of Chicago Press, Chicago, 1941).
[37] http://www.imdb.com/
[38] If one considers the worldwide web to be a social network [an issue of some debate—see B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite, Annu. Rev. Sociol. **22**, 213 (1996)], then it certainly dwarfs the networks studied here, having, it is estimated, about a billion vertices at the time of writing.
[39] P. Hoffman, *The Man Who Loved Only Numbers* (Hyperion, New York, 1998).
[40] P. Erdős and M. Kac, Am. J. Math. **26**, 738 (1940); R.M. Ziff, G.E. Uhlenbeck, and M. Kac, Phys. Rep. **32**, 169 (1977); M.E.J. Newman and R.M. Ziff, Phys. Rev. Lett. **85**, 4104 (2000).
[41] J.W. Grossman and P.D.F. Ion, Congressus Numerantium **108**, 129 (1995).
[42] R. De Castro and J.W. Grossman, Math. Intelligencer **21**, 51 (1999).
[43] V. Batagelj and A. Mrvar, Soc. Networks **22**, 173 (2000).
[44] L. Egghe and R. Rousseau, *Introduction to Informetrics* (Elsevier, Amsterdam, 1990).
[45] O. Persson and M. Beckmann, Scientometrics **33**, 351 (1995).
[46] G. Melin and O. Persson, Scientometrics **36**, 363 (1996).
[47] H. Kretschmer, Z. Sozialpsychol. **29**, 307 (1998).
[48] Y. Ding, S. Foo, and G. Chowdhury, Int. Inf. Lib. Rev. **30**, 367 (1999).
[49] There has been a considerable amount of work on networks of citations between papers, both in information science [see D.J. de Solla Price, Science **149**, 510 (1965), for instance], and more recently in physics [see S. Redner, Eur. Phys. J. B **4**, 131 (1998)]. In these networks, the actors are papers and the (directed) ties between them are citations of one paper by another. However, while citation data are plentiful and many results are known, citation networks are not true social networks since the authors of two papers need not be acquainted for one of them to cite the other's work. On the other hand, citation probably does imply a certain congruence in the subject matter of the two papers, which, although not a social relationship, may certainly be of interest for other reasons.
[50] M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[51] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, e-print cond-mat/0104162.
[52] M.E.J. Newman, e-print cond-mat/0104209.
[53] H.B. O'Connell, e-print physics/0007040.
[54] A.J. Lotka, J. Wash. Acad. Sci. **16**, 317 (1926).
[55] H. Voos, J. Am. Soc. Inf. Sci. **25**, 270 (1974).
[56] M.L. Pao, J. Am. Soc. Inf. Sci. **37**, 26 (1986).
[57] S. Redner (private communication).
[58] A.-L. Barabási (private communication).
[59] P. Erdős and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[60] B. Bollobás, *Random Graphs* (Academic Press, New York, 1985).
[61] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995); Combinatorics, Prob. Comput. **7**, 295 (1998).
[62] M.E.J. Newman, following paper, Phys. Rev. E **64**, 016132 (2001).