Who is the best connected scientist? A study of scientific coauthorship networks

M. E. J. Newman

Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109. U.S.A.

and

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. U.S.A.

Abstract

Using data from 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 have studied many statistical properties of our networks, including numbers of papers written by authors, numbers of authors per paper, numbers of collaborators that scientists have, typical distance through the network from one scientist to another, and a variety of measures of connectedness within a network, such as closeness and betweenness. We further argue that simple networks such as these cannot capture the variation in the strength of collaborative ties and propose a measure of this strength based on the number of papers coauthored by pairs of scientists, and the number of other scientists with whom they worked on those papers. Using a selection of our results, we suggest a variety of possible ways to answer the question "Who is the best connected scientist?"

1 Introduction

A social network is a set of people or groups each of which has connections of some kind to some or all of the others (Wasserman and Faust, 1994; Scott, 2000). 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 twentieth century sociology. In the last few years, it has become the focus of considerable attention in the applied mathematics and statistical physics communities as well (Strogatz, 2001; Barabási, 2002; Watts, 2003; Newman, 2003).

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 (Mariolis, 1975; Galaskiewicz and Marsden, 1978; Padgett and Ansell, 1993), a school (Rapoport and Horvath, 1961; Fararo and Sunshine, 1964), a religious or ethnic community (Bernard *et al.*, 1988), 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, often 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 kinds of quantitative approaches to network analysis that have been developed in physics and mathematics. 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 a thousand actors. This makes the statistical accuracy of many results poor, a particular difficulty for the large-system-size methods adopted 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 school-children, for instance (Rapoport and Horvath, 1961; Fararo and Sunshine, 1964; Moody, 2001), 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.

In response to these inadequacies, many researchers have turned instead to other, better documented networks, for which reliable statistics can be collected. Examples include the Internet (Faloutsos *et al.*, 1999; Chen *et al.*, 2002), the world wide web (Albert *et al.*, 1999; Broder *et al.*, 2000), email networks (Ebel *et al.*, 2002; Newman *et al.*, 2002a), peer-to-peer networks (Adamic *et al.*, 2001; Ripeanu *et al.*, 2002), power grids (Watts and Strogatz, 1998), telephone call graphs (Abello *et al.*, 1998), and train routes (Sen *et al.*, 2002). 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 built 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 studied in the past include women and the social events they attend (Davis *et al.*, 1941), company directors and the boards of directors on which they sit (Mariolis, 1975; Davis and Greve, 1997), company CEOs and the clubs they frequent (Galaskiewicz and Marsden, 1978), and movie actors and the movies in which they appear (Watts and Strogatz, 1998; Amaral *et al.*, 2000). 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 ascer-

tained from membership lists, making time-consuming interviews or questionnaires unnecessary. A network of movie actors, for example, has been compiled using the resources of the Internet Movie Database,[1] 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 article we study in detail another affiliation network, one which is a true social network, for which excellent data are available, and which furthermore will be of interest to readers for personal as well as scientific reasons. In this article, we study networks in which the actors are scientists and the ties between them are scientific collaborations, as documented in the papers that they write.

2 Coauthorship networks

Here we construct networks of scientists in which a link between two scientists is established by their coauthorship of one or more scientific papers. These networks are affiliation networks in which actors are linked by their common membership of groups consisting of the authors of a paper. They are more truly social networks than many affiliation networks; it is probably fair to say that most people who have written a paper together are genuinely acquainted with one another, in a way that, for example, 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 discuss these where appropriate. By and large, however, the network reflects genuine professional interaction between scientists, and may be the largest social network ever studied.[2]

The idea of constructing a network of coauthorship is not new. Many readers will be familiar with the concept of the Erdős number, named for Paul Erdős, the Hungarian mathematician, one of the founding fathers of graph theory, among other things (Hoffman, 1998). 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 a 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 (Erdős and Kac, 1940; Ziff *et al.*, 1977; Newman and Ziff, 2000). In the jargon of social networks, your Erdős number is the geodesic distance between you and Erdős in the coauthorship network. In recent studies (Batagelj and Mrvar, 2000; Grossman, 2002), 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,

[1] `http://www.imdb.com/`.

[2] If one considers the world wide web to be a social network (an issue of some debate—see Wellman *et al.* (1996)), then it certainly dwarfs the networks studied here, with more than three billion pages cataloged by the largest search engines at the time of writing.

results are in most cases found if the network is centered on another mathematician. (On the other hand, fifth-most published mathematician, Lucien Godeaux, produced 644 papers, on 643 of which he was the sole author. He has no finite Erdős number (Grossman and Ion, 1995). 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 (Egghe and Rousseau, 1990; Kretschmer, 1994; Persson and Beckmann, 1995; Melin and Persson, 1996; Ding *et al.*, 1999; Bordens and Gómez, 2000). However, these studies have not so far attempted to reconstruct actual collaboration networks from bibliographic data, concentrating more on organizational and institutional aspects of collaboration.[3]

In this article, we study networks of scientists using bibliographic data drawn from four publicly available databases of papers. The databases are:

1. Physics 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. 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 also professionally maintained. High energy physics is an interesting case socially, having a tradition of much larger experimental collaborations than other disciplines.
4. NCSTRL: a database of preprints in computer science, submitted by participating institutions and stretching back about ten years.

We have constructed networks of collaboration for each of these databases separately and analyzed them using a variety of techniques, some standard, some invented for the purpose.

The outline of the article is as follows. In Sec. 3 we discuss some basic statistics, to give a feel for the shape of our networks. Among other things we discuss the

[3]There has been a considerable amount of work on networks of citations between papers, both in information science (Price, 1965; Egghe and Rousseau, 1990; Seglen, 1992) and more recently in physics (Redner, 1998). These networks, though often confused with coauthorship networks, are quite distinct from them; in a citation network the "actors" are papers and the (directed) ties between them are citations of one paper by another. While citation data are plentiful and many results are known, citation networks are not really 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.

typical numbers of papers per author, authors per paper, and number of collaborators of scientists in the various disciplines. In Sec. 4 we look at a variety of measures concerned with paths between scientists in the network. In Sec. 5 we extend our networks to include a measure of the strength of collaborative ties between scientists and examine measures of connectedness in these weighted networks. In Sec. 6 we give our conclusions. This article is an updated and extended version of an earlier two-part report (Newman, 2001b,c).

3 Basic results

For this study, we constructed collaboration networks using data from a five-year period from January 1, 1995 to December 31, 1999, 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 Physics E-print Archive and the NCSTRL database is relatively poor before 1995, and this sets a limit on how far back we can look. Third, the networks change over time, both because people enter and leave the professions they represent and because practices of scientific collaboration and publishing change. In this article we do not address time evolution of the network (though this is done elsewhere—see Newman (2001d) and Barabási *et al.* (2002)). 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 computer program implementing this procedure, in which names were stored in a simple array, would take time $O(pn)$ to run to completion, 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 1 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.

3.1 Number of authors

The size of the databases varies considerably, from over 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 necessarily the same as the number of authors. A

		Physics E-print Archive					
	Medline	complete	astro-ph	cond-mat	hep-th	SPIRES	NCSTRL
total papers	2163923	98502	22029	22016	19085	66652	13169
total 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 C	0.066(7)	0.43(1)	0.414(6)	0.348(6)	0.327(2)	0.726(8)	0.496(6)
mean distance	4.6(2)	5.9(2)	4.66(7)	6.4(1)	6.91(6)	4.0(1)	9.7(4)
maximum distance	24	20	14	18	19	19	31

Table 1: Summary of results of the analysis of seven scientific collaboration networks. Numbers in parentheses give an estimate of the error on the least significant figures.

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 (1995) point out that there are two American mathematicians named Norman Lloyd Johnson, who are known to be distinct people and who work in different fields, but between whom computer programs such as ours cannot hope to distinguish. Even additional clues such as home institution or field of specialization cannot reliably 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 University of Michigan and the Santa Fe Institute, and publishes in statistical physics, sociology, and epidemiology.

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 1 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.

3.2 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 which contains both preprints 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 (O'Connell, 2000), 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 showed, using a dataset compiled by hand, that this distribution followed a power law, with exponent approximately -2, a result which is now referred to as Lotka's Law of Scientific Productivity (Lotka, 1926). 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

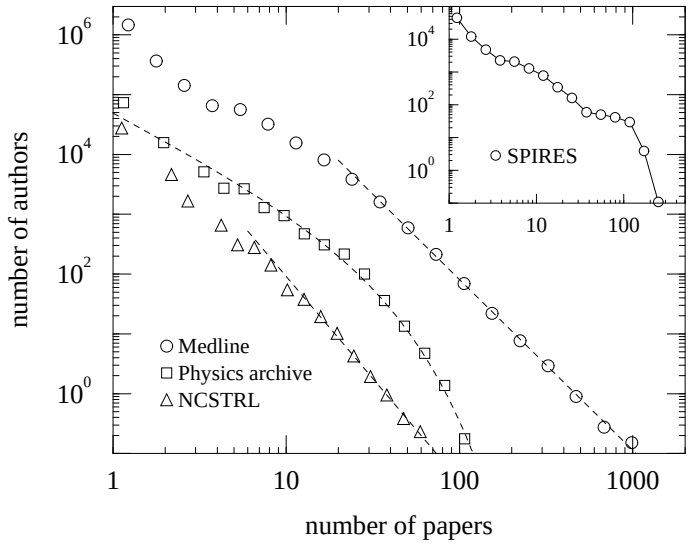

Figure 1: Histograms of the number of papers written by authors in Medline (circles), the physics archive (squares), and NCSTRL (triangles). The dotted lines are fits to the data as described in the text. Inset: the equivalent histogram for the SPIRES database.

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 (Voos, 1974; Pao, 1986; Egghe and Rousseau, 1990). For the physics archive the pure power law is a poor fit. An exponentially truncated power law does much better:

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

where τ and κ are constants and C is fixed by the requirement of normalization—see Fig. 1. (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 a brief inspection of the names of the most published authors. The top ten, for example, are Suzuki, T., Wang, Y., Suzuki, K., Takahashi, M., Nakamura, T., Tanaka, K., Tanaka, T., Wang, J., Suzuki, Y., and Takahashi, T. The predominance of Japanese names in this list may reflect differences in author attribution practices in Japanese biomedical research, but more probably these are simply common names, and these apparently highly published authors are each several different people who have been

conflated in our analysis. Thus it is possible that there is not after all any fat tail in the distribution for the Medline database, only the illusion of one produced by the large number of scientists with commonly occurring names. (This doesn't however explain why the tail appears to follow a power law.) This argument is strengthened by the sheer numbers of papers involved. T. Suzuki published, it appears, 1697 papers, or about one paper a day, including weekends and holidays, every day for the entire five-year course of our study. This seems to be an improbably large output.

Interestingly, no national bias is seen in any of the other databases, and the names that top the list in physics and computer science are not common ones. (For example, the most published authors in the other three databases are Shelah, S. (physics archive), Wolf, G. (SPIRES), and Bestavros, A. (NCSTRL).) Thus it is still unclear why the NCSTRL database should have a power-law tail, though 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 dataset.

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 a hundred papers, the statistics in the tail of the distribution can be substantially skewed by such practices.

In the first column of Table 2, we list the most frequent authors in three subject-specific subdivisions of the physics archive: `astro-ph` (astro-physics), `cond-mat` (condensed matter physics), and `hep-th` (high-energy theory). Although there is only space to list the top ten winners in this table, the entire list (and the corresponding lists for the other tables in this article) can be found by the curious reader on the world wide web.[4]

3.3 Numbers of authors per paper

Grossman and Ion (1995) report that the average number of authors on papers in mathematics has increased steadily over the last sixty years, from a little over 1 to its current value of about 1.5. As Table 1 shows, still higher numbers 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 in our calculations. For databases covering experimental or partly experimental subject areas the averages are 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 how small it is. The hundreds strong mega-collaborations of CERN and Fermilab are sufficiently diluted by theoretical and smaller experimental groups, that the number is only 9, and not 90.)

[4]Complete tables of results for authors in the Physics E-print Archive can be found on the world wide web at `http://www.santafe.edu/~mark/collaboration/`.

	number of papers		number of co-workers		betweenness		collaboration weight	
astro-ph	112	Fabian, A.C.	360	Frontera, F.	2.33	Kouveliotou, C.	16.5	Moskalenko, I.V./Strong, A.W.
	101	van Paradijs, J.	353	Kouveliotou, C.	2.15	van Paradijs, J.	15.0	Hernquist, L./Heyl, J.S.
	81	Frontera, F.	329	van Paradijs, J.	1.80	Filippenko, A.V.	14.0	Mathews, W.G./Brighenti, F.
	80	Hernquist, L.	299	Piro, L.	1.57	Beaulieu, J.P.	13.4	Labini, F.S./Pietronero, L.
	79	Gould, A.	296	Costa, E.	1.52	Nomoto, K.	12.2	Piran, T./Sari, R.
	78	Silk, J.	291	Feroci, M.	1.52	Pian, E.	11.8	Zaldarriaga, M./Seljak, U.
	78	Klis, M.V.D.	284	Pian, E.	1.49	Frontera, F.	11.4	Hernquist, L./Katz, N.
	73	Kouveliotou, C.	284	Hurley, K.	1.35	Silk, J.	11.1	Avila-Reese, V./Firmani, C.
	70	Ghisellini, G.	244	Palazzi, E.	1.33	Kamionkowski, M.	10.9	Dai, Z.G./Lu, T.
	66	Piro, L.	244	Heise, J.	1.28	McMahon, R.G.	10.8	Ostriker, J.P./Cen, R.
cond-mat	116	Parisi, G.	107	Uchida, S.	4.11	MacDonald, A.H.	22.3	Belitz, D./Kirkpatrick, T.R.
	79	Scheffler, M.	103	Ueda, Y.	3.96	Bishop, A.R.	17.0	Shrock, R./Tsai, S.
	75	Das Sarma, S.	96	Revcolevschi, A.	3.36	Das Sarma, S.	15.0	Yukalov, V.I./Yukalova, E.P.
	74	Stanley, H.E.	94	Eisaki, H.	2.96	Tosatti, E.	14.7	Martín-Delgado, M.A./Sierra, G.
	70	MacDonald, A.H.	84	Cheong, S.	2.52	Wang, X.	14.3	Krapivsky, P.L./Ben-Naim, E.
	68	Sornette, D.	83	Isobe, M.	2.38	Revcolevschi, A.	14.1	Beenakker, C.W.J./Brouwer, P.W.
	60	Volovik, G.E.	78	Stanley, H.E.	2.30	Uchida, S.	13.8	Weng, Z.Y./Sheng, D.N.
	56	Beenakker, C.W.J.	76	Shirane, G.	2.21	Sigrist, M.	13.7	Sornette, D./Johansen, A.
	53	Dagotto, E.	76	Scheffler, M.	2.19	Cheong, S.	13.6	Rikvold, P.A./Novotny, M.A.
	50	Helbing, D.	76	Menovsky, A.A.	2.18	Stanley, H.E.	13.0	Scalapino, D.J./White, S.R.
hep-th	78	Odintsov, S.D.	50	Ambjorn, J.	0.98	Odintsov, S.D.	34.0	Lu, H./Pope, C.N.
	73	Lu, H.	44	Ferrara, S.	0.88	Ambjorn, J.	29.0	Odintsov, S.D./Nojiri, S.
	72	Pope, C.N.	43	Vafa, C.	0.88	Kogan, I.I.	18.7	Lee, H.W./Myung, Y.S.
	69	Cvetic, M.	39	Odintsov, S.D.	0.84	Henneaux, M.	18.3	Schweigert, C./Fuchs, J.
	68	Ferrara, S.	39	Kogan, I.I.	0.73	Douglas, M.R.	14.7	Ovrut, B.A./Waldram, D.
	65	Vafa, C.	36	Proeyen, A.V.	0.67	Ferrara, S.	14.7	Kleihaus, B./Kunz, J.
	65	Tseytlin, A.A.	35	Fre, P.	0.63	Vafa, C.	12.9	Mavromatos, N.E./Ellis, J.
	65	Mavromatos, N.E.	35	Ellis, J.	0.60	Khare, A.	12.4	Kachru, S./Silverstein, E.
	63	Witten, E.	35	Douglas, M.R.	0.58	Tseytlin, A.A.	11.7	Kakushadze, Z./Tye, S.H.H.
	54	Townsend, P.K.	34	Lu, H.	0.58	Townsend, P.K.	11.6	Arefeva, I.Y./Volovich, I.V.

Table 2: The authors with the highest numbers of papers, numbers of coauthors, and betweenness, and strongest collaborations in astrophysics, condensed matter physics, and high-energy theory. The figures for betweenness have been divided by 10^6. Full lists of the rankings of all the authors in these databases can be found on the world wide web at `http://www.santafe.edu/~mark/collaboration/`.

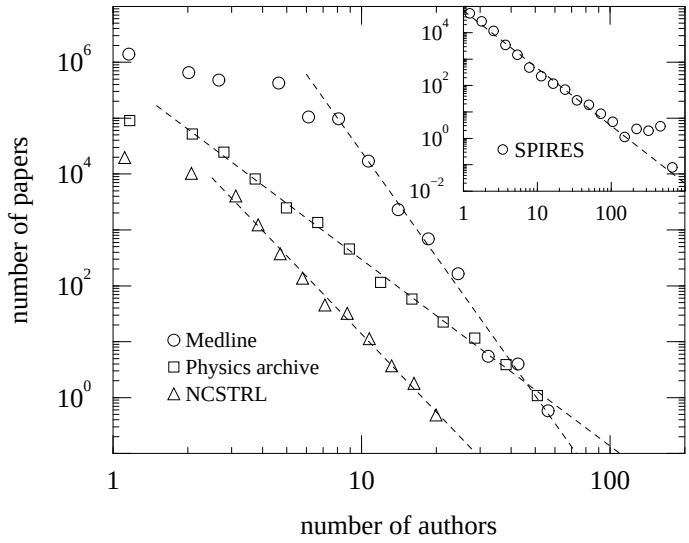

Figure 2: Histograms of the number of authors on papers in Medline (circles), the physics archive (squares), and NCSTRL (triangles). The dotted lines are the best fit power-law forms. Inset: the equivalent histogram for the SPIRES database, showing a clear peak in the 200 to 500 author range.

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), $-3.34(5)$ (physics archive), $-4.6(1)$ (NCSTRL), and $-2.18(7)$ (SPIRES). The SPIRES data, which are again shown in a separate inset, display a pronounced peak in the distribution around 200–500 authors. This peak presumably corresponds to the large experimental collaborations which dominate the upper end of this histogram.

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

3.4 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 social network of the high-energy physics community; it seems unlikely that an author could know 173 colleagues well.

The distributions of numbers of collaborators are shown in Fig. 3. In all cases

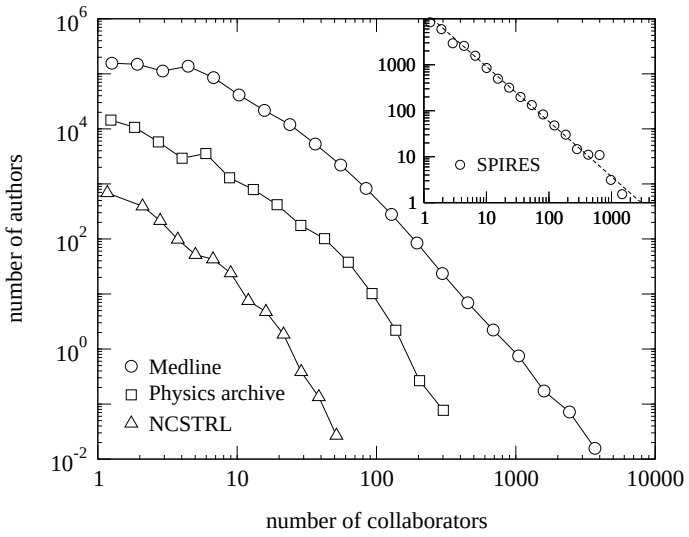

Figure 3: Histograms of the number of collaborators of authors in Medline (circles), the physics archive (squares), and NCSTRL (triangles). Inset: the equivalent histogram for the SPIRES database, which is well fit by a power law (dotted line).

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 product of the predominance of large collaborations.

For the other three databases, the distributions show some curvature. This may, as we have previously suggested, be the signature of an exponential cutoff, produced once again by the finite time window of the study (Newman, 2001a). Redner (personal communication) and Barabási *et al.* (2002) have independently suggested alternative explanations based on growth models of networks, although the fundamental causative agent is the same finite time window in these theories also.

Column 2 of Table 2 shows the authors in `astro-ph`, `cond-mat`, and `hep-th` with the largest numbers of collaborators. The winners in this race tend to be experimentalists, who conduct research within larger collaborations, although there are exceptions. The high-energy theory database of course contains only theorists, and the smaller numbers of collaborators reflect this.

3.5 Size of the giant component

In the theory of random graphs (Erdős and Rényi, 1960; Bollobás, 2001) 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 usually fills a large portion of the graph, and all other components (i.e., connected subsets of vertices) are small with mean size independent of the size of the network. 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 1 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 a large extent with the longer established SPIRES database, and it is possible that some authors neglect it for this reason (O'Connell, 2000). 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.

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. 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.

3.6 Clustering coefficients

An interesting idea circulating in social network theory currently is that of **transitivity**, which, along with its sibling **structural balance**, describes symmetry of interaction amongst 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 (Watts and Strogatz, 1998), and can be written[5]

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

The factor of three 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 unipartite random graphs $C = \mathrm{O}(n^{-1})$, where n is the number of vertices, and hence goes to zero in the limit of large graph size (Watts and Strogatz, 1998; Newman, 2003). In social networks it is believed that the clustering coefficient will take a non-zero value even in very large networks, because there is a finite (and probably quite large)

[5]This is one of two slightly different definitions of the clustering coefficient that are in use. See, for instance, Newman (2003).

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 1 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 which mimics the trivial clustering effect, and the results indicate that only about a half of the clustering we see is a result of authors collaborating in groups of three or more (Newman *et al.*, 2001). 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 (Wasserman and Faust, 1994).

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 person, 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 who specialize in them.

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 postdocs, students, and technicians working on different projects. This organization produces a tree-like hierarchy of collaborative ties with fewer interactions within levels of the tree than between them. 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 tree-like 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.

4 Distances and centrality

The basic statistics of the previous section are certainly of importance, particularly for the construction of network models (Watts and Strogatz, 1998; Albert *et al.*, 1999; Kleinberg, 2000; Newman *et al.*, 2001; Krapivsky and Redner, 2001), but there is much more that we can do with our collaboration networks. In this section, we look at some simple but useful measures of network structure, concentrating on measures having to do with paths between vertices in the network. In Sec. 5 we discuss some shortcomings of these measures, and construct some new and more complex measures that may better reflect true collaboration patterns.

4.1 Shortest paths

A fundamental concept in graph theory is the **geodesic**, the shortest path of vertices and edges that links two given vertices. There may not be a unique geodesic between two vertices: there may be two or more shortest paths, which may or may not share some vertices. Or there may be no paths between the vertices at all. The geodesic(s) between two vertices s and t can be calculated in time $O(m)$, where m is the number of edges in the graph, using the following algorithm, which is a modified form of the standard breadth-first search (Cormen *et al.*, 2001).

1. Assign vertex s distance zero, to indicate that it is zero steps away from itself, and set $d = 0$.

2. For each vertex i whose assigned distance is d, follow each attached edge to the vertex j at its other end and then do one of the following three things:

   (a) If j has not already been assigned a distance, assign it distance $d + 1$. Declare i to be a predecessor of j.

   (b) If j has already been assigned distance $d + 1$, then there is no need to do this again, but i is still declared a predecessor of j.

   (c) If j has already been assigned a distance less than $d + 1$, do nothing.

3. Set $d \leftarrow d + 1$.

4. Repeat from step (2) until there are no unassigned vertices left.

Now the shortest path (if there is one) between s and t is the path you get by stepping from t to its predecessor, and then to the predecessor of each successive vertex until s is reached. If a vertex has two or more predecessors, then there are two or more shortest paths, each of which must be followed separately if we wish to know all shortest paths between s and t.

In Fig. 4 we show the shortest paths of known collaborations between two of the author's colleagues, Duncan Watts (Columbia) and László Barabási (Notre Dame), both of whom work on networks of various kinds. It is interesting to note that,

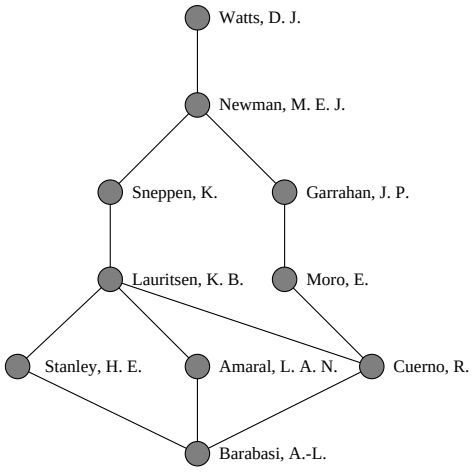

Figure 4: The geodesics, or shortest paths, in the collaboration network of physicists between Duncan Watts and László Barabási.

although the two scientists in question are well acquainted both personally and with one another's work, the shortest path between them does not run entirely through other collaborations in the field. (For example, the connection between the present author and Juan Pedro Garrahan results from our coauthorship of a paper on spin glasses.) Although this may at first sight appear odd, it is probably in fact a good sign. It indicates that workers in the field come from different scientific camps, rather than all descending intellectually from a single group or institution. This presumably increases the likelihood that those workers will express independent opinions on the open questions of the field, rather than merely spouting slight variations on the same underlying doctrine.

A database that would allow one conveniently and quickly to extract shortest paths between scientists in this way might have some practical use. Kautz *et al.* (1997) have constructed a web-based system which does just this for computer scientists, with the idea that such a system might help to create new professional contacts by providing a "referral chain" of intermediate scientists through whom contact may be established.

4.2 Betweenness and funneling

A quantity of interest in many social network studies is the **betweenness** of an actor i, which is defined as the total number of shortest paths between pairs of actors that pass through i (Freeman, 1977). This quantity is one possible indicator of who the most influential people in the network are. In a network in which information flows entirely or mostly along the shortest paths between actors, those with highest betweenness are the ones who control the flow of information between most others. The vertices with highest betweenness also produce an increase in the geodesic distance between the largest number of pairs of others when removed from the network (Wasserman and Faust, 1994).

Naively, one might think that betweenness would take time of order $O(mn^2)$ to

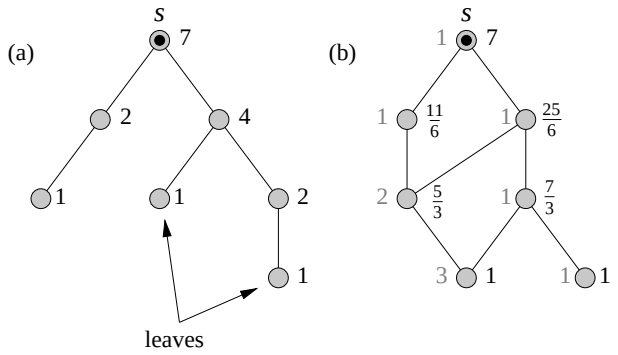

Figure 5: Calculation of betweenness: (a) When there is only a single shortest path from a source vertex s (top) to all other reachable vertices, those paths necessarily form a tree, which makes the calculation of the contribution to betweenness from this set of paths particularly simple, as described in the text. (b) For cases in which there is more than one shortest path to some vertices, the calculation is more complex. First we must calculate the number of paths from the source s to each other vertex (numbers to left of vertices), and then use these to weight the path counts appropriately and derive the betweenness scores (numbers to right of vertices).

calculate for all vertices, since there are $O(n^2)$ shortest paths to be considered, each of which takes time $O(m)$ to calculate, and until recently the standard network analysis packages such as UCInet and Pajek indeed used $O(mn^2)$ algorithms. Recently however, faster algorithms for betweenness have been discovered by the present author (Newman, 2001c) and independently by Brandes (2001). These algorithms can perform the same calculation in time $O(mn)$. Here we describe the algorithm of Newman (2001c), which is fast enough to allow the exhaustive calculation of betweenness for all vertices in the very large graphs studied here.

We start by performing a breadth-first search to determine the set of shortest paths from some source vertex s to all other vertices that are reachable from s. Consider first the simple case of a network in which there is only a single shortest path from the source vertex to any other. (We will consider other cases in a moment.) The resulting set of paths then forms a tree as shown in Fig. 5a. We can use this tree to calculate betweenness as follows. We find first the "leaves" of the tree, i.e., those nodes such that no shortest paths to other nodes pass through them, and we assign a score of 1 to them—the only path to these vertices is the one that ends there. Then, starting with those vertices that are farthest from the source vertex s on the tree, i.e., lowest in Fig. 5a, we work upwards, assigning a score to each vertex that is 1 plus the sum of the scores on the neighboring vertices immediately below it. When we have gone though all vertices in the tree, the resulting scores are the betweenness counts for the paths from vertex s. (In our calculation we define paths to include the vertices at their ends. Sometimes they are defined to exclude these vertices, in which case the score at each vertex is decreased by 1, except for the source vertex s, which receives a score of zero.) Repeating the process for all possible vertices s and summing the scores, we arrive at the full betweenness scores for shortest paths between all pairs. The breadth-first search and the process of working up through the tree both take

worst-case time $O(m)$ and there are n vertices total, so the entire calculation takes time $O(mn)$ as claimed.

This simple case serves to illustrate the basic principle behind the algorithm. In general, however, it is not the case that there is only a single shortest path between any pair of vertices. Most networks have at least some vertex pairs between which there are several geodesic paths of equal length. Figure 5b shows a simple example of a shortest path "tree" for a network with this property. The resulting structure is in fact no longer a tree, and in such cases an extra step is required in the algorithm to correctly calculate the betweenness.

Following Freeman's original definition of betweenness (Freeman, 1977), we give multiple shortest paths between a pair of vertices equal weights summing to 1. Note that some of the paths may run through the same vertices for some part of their length, resulting in vertices with greater weight. To calculate correctly what fraction of the paths flow through each vertex in the network, we generalize the breadth-first search part of our algorithm, as follows.

Consider Fig. 5b and suppose we are starting at vertex s. We carry out the following steps:

1. Assign vertex s distance zero, to indicate that it is zero steps from itself, and set $d = 0$. Also assign s a weight $w_s = 1$ (whose purpose will become clear shortly).

2. For each vertex i whose assigned distance is d, follow each attached edge to the vertex j at its other end and then do one of the following three things:

   (a) If j has not yet been assigned a distance, assign it distance $d+1$ and weight $w_j = w_i$.

   (b) If j has already been assigned a distance and that distance is equal to $d+1$, then the vertex's weight is increased by w_i, that is $w_j \leftarrow w_j + w_i$.

   (c) If j has already been assigned a distance less than $d+1$, do nothing.

3. Set $d \leftarrow d + 1$.

4. Repeat from step 2 until there are no vertices that have distance d.

The resulting weights for the example of Fig. 5b are shown to the left of each vertex in the figure.

Physically, the weight on a vertex i represents the number of distinct paths from the source vertex to i. These weights are precisely what we need to calculate our betweennesses, because if two vertices i and j are connected, with j farther than i from the source s, then the fraction of a geodesic path from j through i to s is given by w_i/w_j. Thus, to calculate the contribution to the betweenness from all shortest paths starting at s, we need only carry out the following steps:

1. Find every "leaf" vertex t, i.e., a vertex such that no paths from s to other vertices go though t and assign it a score of $x_t = 1$.

2. Now, starting with the vertices that are farthest from the source vertex s—lower down in a diagram such as Fig. 5b—work up towards s. To each vertex i assign a score $x_i = 1 + \sum_j x_j w_i/w_j$, where the sum is over the neighbors j immediately below vertex i.

3. Repeat from step 2 until vertex s is reached.

The resulting scores are shown to the right of each vertex in Fig. 5b. Now repeating this process for all n source vertices s and summing the resulting scores on the vertices gives us the total betweenness for all vertices in time $O(mn)$.

We have applied this algorithm to our coauthorship networks and in column 3 of Table 2 we show the ten highest betweennesses in the `astro-ph`, `cond-mat`, and `hep-th` subdivisions of the physics archive. While we leave it to the knowledgeable reader to decide whether the scientists named are indeed pivotal figures in their respective fields, we do notice one interesting feature of the results. The betweenness measure gives very clear winners in the competition: the individuals with highest betweenness are well ahead of those with second highest, who are in turn well ahead of those with third highest, and so on. This same phenomenon has been noted in other networks (Wasserman and Faust, 1994; Goh *et al.*, 2001).

Strogatz has raised an interesting question about social networks which we can address using our betweenness algorithm: are all of your collaborators equally important for your connection to the rest of the world, or do most paths from others to you pass through just a few of your collaborators (S. H. Strogatz, personal communication)? One could certainly imagine that the latter might be true. Collaboration with just one or two senior or famous members of one's field could easily establish short paths to a large part of the collaboration network, and all of those short paths would go through those one or two members. Strogatz calls this effect "funneling." Since our algorithm, as a part of its operation, calculates the vertices through which each geodesic path to a specified actor passes, it is a trivial modification to calculate also how many of those geodesic paths pass through each of the immediate collaborators of that actor, and hence to use it to look for funneling.

Our collaboration networks, it turns out, show strong funneling. For most people, their top few collaborators lie on most of the paths between themselves and the rest of the network. The rest of their collaborators, no matter how numerous, account for only a small number of paths. Consider, for example, the present author. Out of the 44 000 scientists in the giant component of the physics archive collaboration network, 31 000 paths from them to me, about 70%, pass through just two of my collaborators, Chris Henley and Juanpe Garrahan. Another 13 000, most of the remainder, pass through the next four collaborators. The remaining five account for a mere 1% of the total.

To give a more quantitative impression of the funneling effect, we show in Fig. 6 the average fraction of paths that pass through the top 10 collaborators of an author, averaged over all authors in the giant component of the Physics database. The figure shows for example that on average 64% of one's shortest paths to other scientists pass through one's top-ranked collaborator. Another 17% pass through the second-ranked collaborator. The top 10 shown in the figure account for 98% of all paths.

That one's top few acquaintances account for most of one's shortest paths to the rest of the world has been noted before in other contexts. For example, Stanley Milgram, in his famous "small-world" experiment (Milgram, 1967), noted that most of the paths he found to a particular target person in an acquaintance network went through just one or two acquaintances of the target. He called these people "sociometric superstars."

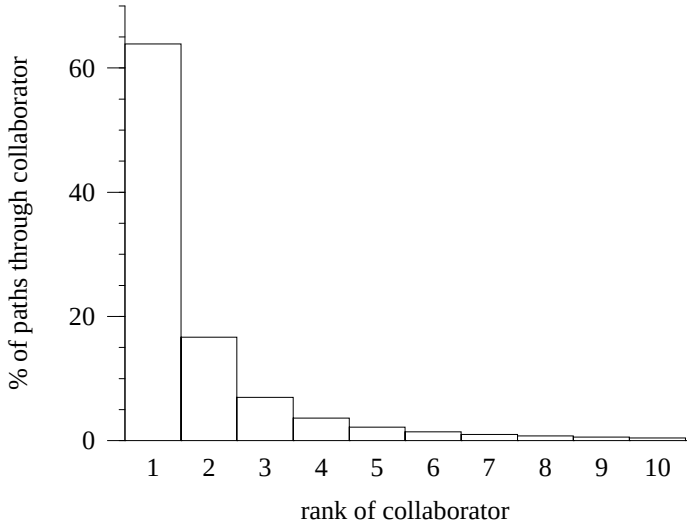

Figure 6: The average percentage of paths from other scientists to a given scientist that pass through each collaborator of that scientist, ranked in decreasing order. The plot is for the physics archive network, although similar results are found for other networks.

4.3 Average distances

Breadth-first search allows to us calculate exhaustively the lengths of the shortest paths from every vertex on a graph to every other in time $O(mn)$. We have done this for each of the networks studied here and averaged these distances to find the average distance between any pair of (connected) authors in each of the subject fields studied. These figures are given in the penultimate row of Table 1. As the table shows, these figures are all quite small: they vary from 4.0 for SPIRES to 9.7 for NCSTRL, although this last figure may be artificially inflated by the poor coverage of this database discussed in Sec. 3.5. At any rate, all the figures are very small compared to the number of vertices in the corresponding databases. This "small-world effect," famously discussed by Milgram (1967) and by Pool and Kochen (1978), is, like the existence of the giant component, probably a good sign for science; it shows that scientific information—discoveries, experimental results, theories—will not have far to travel through the network of scientific acquaintance to reach the ears of those who can benefit by it. Even the *maximum* distances between scientists in these networks, shown in the last row of the table, are not very large, the longest path in any of the networks being just 31 steps long, again in the NCSTRL database.

The explanation of the small-world effect is simple. Consider Fig. 7, which shows all the collaborators of the present author (in all subjects, not just physics), and all the collaborators of those collaborators—all my first and second neighbors in the collaboration network. As the figure shows, I have 35 first neighbors, but 891 second neighbors. The "radius" of the whole network around me is reached when the number of neighbors within that radius equals the number of scientists in the giant component of the network, and if the increase in numbers of neighbors with distance continues

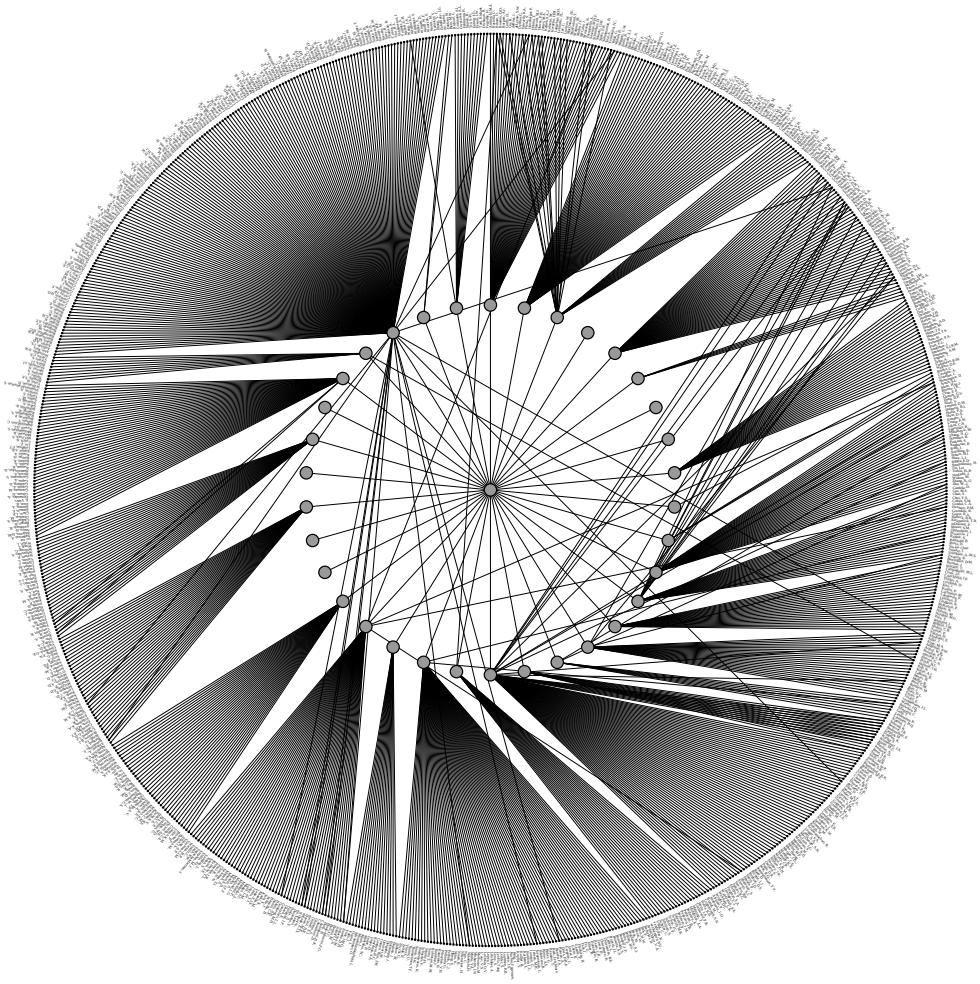

Figure 7: The point in the center of the figure represents the author of the article you are reading, the first ring his collaborators, and the second ring their collaborators. Collaborative ties between members of the same ring, of which there are many, have been omitted from the figure for clarity.

at the impressive rate shown in the figure, it will not take many steps to reach this point.

This simple idea is borne out by theory. In almost all networks, the average distance between pairs of vertices ℓ scales logarithmically with the number of vertices n. In a standard random graph (Erdős and Rényi, 1960; Bollobás, 2001), for instance, $\ell = \log n / \log z$, where z is the average degree of a vertex, the average number of collaborators in our terminology. In the more general class of random graphs in which the distribution of vertex degrees is arbitrary (Bollobás, 1980; Łuczak, 1992; Molloy and Reed, 1995, 1998), rather than Poissonian as in the standard case, the equivalent

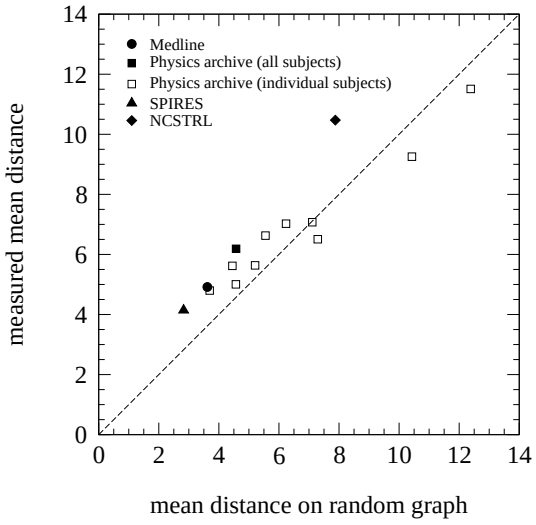

Figure 8: Average distance between pairs of scientists in the various networks, plotted against average distance on a random graph of the same size and degree distribution. The dotted line shows where the points would fall if measured and predicted results agreed perfectly. The solid line is the best straight-line fit to the data.

expression is

$$\ell = \frac{\log(n/z_1)}{\log(z_2/z_1)} + 1, \tag{3}$$

where z_1 and z_2 are the average numbers of first and second neighbors of a vertex (Newman *et al.*, 2001, 2002b). It is widely assumed that this logarithmic behavior extends to most networks, so the small-world effect is not a surprise to those familiar with graph theory. However, it would be nice to demonstrate explicitly the presence of logarithmic scaling in our networks. Figure 8 does this in a crude fashion. In this figure we have plotted the measured value of ℓ, as given in Table 1, against the value given by Eq. (3) for each of our four databases, along with separate points for nine of the subject-specific subdivisions of the physics archive. As the figure shows, the correlation between measured and predicted values is quite good. The correlation coefficient is $R^2 = 0.86$, rising to $R^2 = 0.95$ if the NCSTRL database, with its incomplete coverage, is excluded (the diamond in the figure).

Figure 8 needs to be taken with a pinch of salt. Its construction implicitly assumes that the different networks are statistically similar to one another and to the random graphs with the same distributions of vertex degree, an assumption which is almost certainly not correct. Nonetheless, the fact that even with such inherent errors the logarithmic behavior is still clearly visible lends at least some credence to its graph theoretical basis.

We can also trivially use our breadth-first search algorithm to calculate the average distance from a single vertex to all other vertices in the giant component. This average is essentially the same as the quantity known as **closeness centrality** to

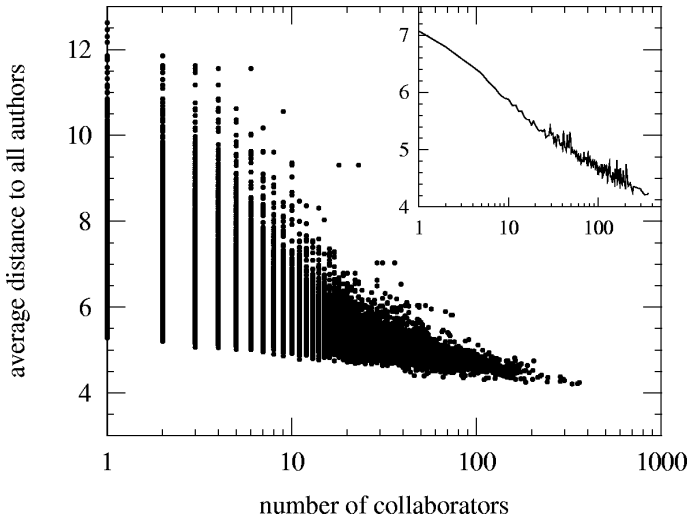

Figure 9: Scatter plot of the mean distance from each physicist in the giant component of the physics archive network to all others as a function of number of collaborators. Inset: the same data averaged vertically over all authors having the same number of collaborators.

social network analysts.[6] Like betweenness it is also a measure, in some sense, of the centrality of a vertex—authors with low values of this average will, it is assumed, be the first to learn new information, and information originating with them will reach others quicker than information originating with other sources. Average distance is thus a measure of centrality of an actor in terms of their access to information, whereas betweenness is a measure of an actor's control over information flowing between others.

Calculating average distance for many networks returns results which look sensible to the observer. Calculations for the network of collaborations between movie actors, for instance, give small average distances for actors who are famous—ones many of us will have heard of. Interestingly, however, performing the same calculation for our scientific collaboration networks does not give exactly the results we might expect. For example, one finds that the people at the top of the list are always experimentalists. This, you might think, is not such a bad thing: perhaps the experimentalists are better connected people? In a sense, in fact, it turns out that they are. In Fig. 9 we show the average distance from scientists in the physics archive to all others in the giant component as a function of their number of collaborators. As the figure shows, there is a clear trend towards shorter average distance as the number of collaborators becomes large. This trend is clearer still in the inset, where we show the same data averaged over all authors who have the same number of collaborators. Since experimentalists often work in large groups, it is not surprising to learn that they tend to have shorter average distances to other scientists.

But this brings up an interesting question, one that we touched upon in Sec. 2:

[6] Technically, closeness is the reciprocal of the average distance to other vertices (Wasserman and Faust, 1994).

while most pairs of people who have written a paper together will know one another reasonably well, there are exceptions. On a high-energy physics paper with 1000 coauthors, for instance, it is unlikely that every one of the 499 500 possible acquaintanceships between pairs of those authors will actually be realized. Our closeness measure does not take into account the tendency for collaborators in large groups not to know one another, or to know one another less well. In the next section we describe a more sophisticated calculation which does do this.

5 Weighted collaboration networks

There is more information present in the databases used here than in the simple networks we have constructed from them, which tell us only whether scientists have collaborated or not. In particular, we also know on how many papers each pair of scientists collaborated during the period of the study, and how many other coauthors they had on each of those papers.[7] We can use this information to make an estimate of the strength of collaborative ties.

First of all, it is probably the case, as we pointed out at the end of the previous section, that two scientists whose names appear on a paper together with many other coauthors know one another less well on average than two who were the sole authors of a paper. The extreme case which we discussed of a very large collaboration illustrates this point forcefully, but it applies to smaller collaborations too. Even on a paper with four or five authors, the authors probably know one another less well on average than authors on a paper with fewer. To account for this effect, we weight collaborative ties inversely according to the number of coauthors as follows. Suppose a scientist collaborates on the writing of a paper that has n authors in total, i.e., he or she has $n-1$ coauthors on that paper. Then we assume that he or she is acquainted with each coauthor $1/(n-1)$ times as well, on average, as if there were only one coauthor. One can imagine this as meaning that the scientist divides his or her time equally between the $n-1$ coauthors. This is obviously only a rough approximation: in reality a scientist spends more time with some coauthors than with others. However, in the absence of other data, it is the obvious first approximation to make.

Second, authors who have written many papers together will, we assume, know one another better on average than those who have written few papers together. To account for this, we add together the strengths of the ties derived from each of the papers written by a particular pair of individuals. Thus, if δ_i^k is one if scientist i was a coauthor of paper k and zero otherwise, then our weight w_{ij} representing the strength of the collaboration (if any) between scientists i and j is

$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1}, \tag{4}$$

where n_k is the number of coauthors of paper k and we explicitly exclude from our

[7] In fact, the full coauthorship pattern is, like all affiliation networks, most properly represented as a bipartite graph with two kinds of vertices representing scientists and papers, and edges running between scientists and the papers on which their name appears as a coauthor. We have investigated this representation elsewhere (Newman *et al.*, 2001, 2002b).

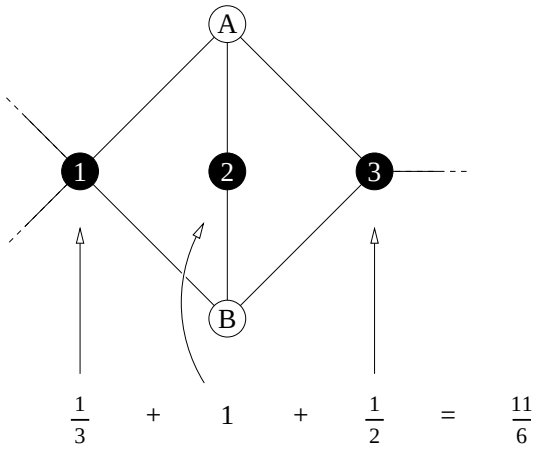

Figure 10: Authors A and B have coauthored three papers together, labeled 1, 2, and 3, which had respectively four, two, and three authors. The tie between A and B accordingly accrues weight $\frac{1}{3}$, 1, and $\frac{1}{2}$ from the three papers, for a total weight of $\frac{11}{6}$.

sums all single-author papers. (They do not contribute to the coauthorship network, and their inclusion in Eq. (4) would make w_{ij} ill-defined.) We illustrate this measure for a simple example in Fig. 10.

Note that the equivalent of vertex degree for our weighted network—i.e., the sum of the weights for each of an individual's collaborations—is now just equal to the number of papers they have coauthored with others:

$$\sum_{j(\neq i)} w_{ij} = \sum_k \sum_{j(\neq i)} \frac{\delta_i^k \delta_j^k}{n_k - 1} = \sum_k \delta_i^k. \tag{5}$$

In Fig. 11 we show as an example collaborations between Gerard Barkema (one of the present author's frequent collaborators) and all of his collaborators in the physics archive for the five years of our study. Lines between points represent collaborations, with their thickness proportional to the weights w_{ij} of Eq. (4). As the figure shows, Barkema has collaborated closely with myself and with Normand Mousseau, and less closely with a number of others. Also, two of his collaborators, John Cardy and Gesualdo Delfino, have collaborated quite closely with one another.

In the last column of Table 2 we show the pairs of collaborators who have the strongest collaborative ties in three subdivisions of the physics archive.

We have used our weighted collaboration graphs to calculate distances between scientists. In this simple calculation we assumed that the distance between authors is just the inverse of the weight of their collaborative tie. Thus if one pair of authors know one another twice as well as another pair, the distance between them is half as great. Calculating minimum distances between vertices on a weighted graph such as this cannot be done using the breadth-first search algorithm of Sec. 4.1, since the shortest weighted path may not be the shortest in terms of number of steps on the unweighted network. Instead we use Dijkstra's algorithm (Ahuja *et al.*, 1993; Cormen *et al.*, 2001), which calculates all distances from a given starting vertex s as follows.

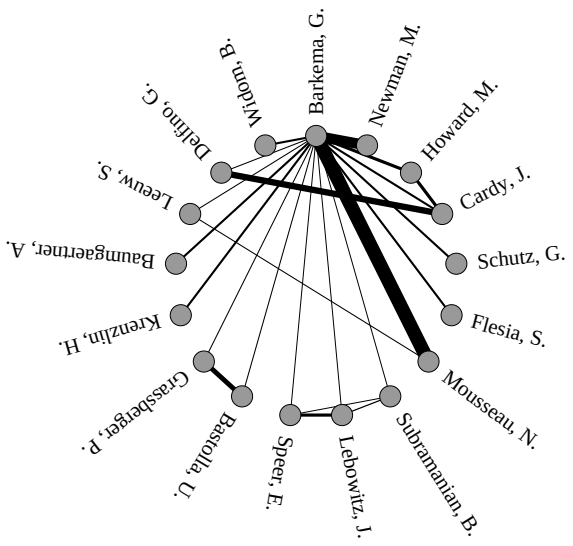

Figure 11: Gerard Barkema and his collaborators, with lines representing collaborations whose thickness is proportional to our estimate, Eq. (4), of the strength of the corresponding tie.

1. Distances from vertex s are stored for each vertex and each distance is labeled either "exact," meaning we have calculated that distance exactly, or "estimated," meaning we have made an estimate of the distance, but that estimate may be wrong. Estimated distances in Dijkstra's algorithm are always upper bounds on the exact distance. We start by assigning an estimated distance of ∞ to all vertices except vertex s to which we assign an estimated distance of zero. (We know the latter to be exactly correct, but for the moment we consider it merely "estimated.")

2. From the set of vertices whose distances from s are currently marked "estimated," choose the one with the lowest estimated distance, and mark this "exact."

3. Calculate the distance from that vertex to each of its immediate neighbors in the network by adding to its distance the length of the edges leading to those neighbors. Any of these distances that is shorter than a current estimated distance for the same vertex supersedes that current value and becomes the new estimated distance for the vertex.

4. Repeat from step (2), until no "estimated" distances remain.

A naive implementation of this algorithm takes time $O(mn)$ to calculate distances from a single vertex to all others, or $O(mn^2)$ to calculate all pairwise distances. One of the factors of n, however, arises because it takes time $O(n)$ to search through

	rank	name	co-workers	papers
astro-ph:	1	Rees, M. J.	31	36
	2	Miralda-Escude, J.	36	34
	3	Fabian, A. C.	156	112
	4	Waxman, E.	15	30
	5	Celotti, A.	119	45
	6	Narayan, R.	65	58
	7	Loeb, A.	33	64
	8	Reynolds, C. S.	45	38
	9	Hernquist, L.	62	80
	10	Gould, A.	76	79
cond-mat:	1	Fisher, M. P. A.	21	35
	2	Balents, L.	24	29
	3	MacDonald, A. H.	64	70
	4	Senthil, T.	9	13
	5	Das Sarma, S.	51	75
	6	Millis, A. J.	43	37
	7	Ioffe, L. B.	16	27
	8	Sachdev, S.	28	44
	9	Lee, P. A.	24	34
	10	Jungwirth, T.	27	17
hep-th:	1	Cvetic, M.	33	69
	2	Behrndt, K.	22	41
	3	Tseytlin, A. A.	22	65
	4	Bergshoeff, E.	21	39
	5	Youm, D.	3	30
	6	Lu, H.	34	73
	7	Klebanov, I. R.	29	47
	8	Townsend, P. K.	31	54
	9	Pope, C. N.	33	72
	10	Larsen, F.	11	27

Table 3: The ten best connected individuals in three of the communities studied here, calculated using the weighted distance measure described in the text.

the vertices to find the one with the smallest estimated distance. The speed of this operation can be improved by storing the estimated distances in a binary heap (a partially ordered binary tree with its smallest entry at its root). We can find the smallest distance in such a heap in time $O(1)$, and add and remove entries in time $O(\log n)$. This speeds up the operation of the algorithm to $O(mn \log n)$, making the calculation feasible for the large networks studied here.

It is in theory possible to generalize any of the calculations of Sec. 4 to the weighted collaboration graph using this algorithm and variations on it. For example, we can find shortest paths between specified pairs of scientists, as a way of establishing referrals. We can calculate the weighted equivalent of betweenness by a simple adaption of our algorithm of Sec. 4.2—we use Dijkstra's algorithm to establish the hierarchy of

predecessors of vertices and then count paths through vertices exactly as before. We can also study the weighted version of the "funneling" effect using the same algorithm. Here we carry out just one calculation explicitly to demonstrate the idea; we calculate the weighted version of the distance centrality measure of Sec. 4.3, i.e., the average weighted distance from a vertex to all others. In Table 3 we show the winners in this particular popularity contest, along with their numbers of collaborators and papers in the database. Many of the scientists who score highly here do indeed appear to be well connected individuals. For example, number 1 best connected astrophysicist, Martin Rees, is the Astronomer Royal of Great Britain.[8] What is interesting to note however (apart from nonchalantly checking to see if one has made it into the top 10) is that sheer number of collaborators is no longer a necessary prerequisite for being well-connected in this sense (although some of the scientists listed do have a large number of collaborators). The case of D. Youm is particularly startling, since Youm has only three collaborators listed in the database but nonetheless is fifth best connected high-energy theorist (out of eight thousand), because those three collaborators are themselves very well connected, and because their ties to Youm are very strong. Experimentalists no longer dominate the field, although the well-connected among them still score highly.

Note that the number of papers for each of the well-connected scientists listed is high. Having written a large number of papers is, as it rightly should be, always a good way of becoming well connected. Whether you write many papers with many different authors, or many with a few, writing many papers will put you in touch with your peers.

6 Conclusions

In this article we have studied social networks of scientists in which the actors are authors of scientific papers, and a tie between two authors 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 cataloged a large number of basic 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 also 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 have also looked at a variety of non-local properties of our networks. We find that typical distances between pairs of authors through the networks are small—the networks form a "small world" in the sense discussed by Milgram—and that they scale logarithmically with total number of authors in a network, in reasonable agreement with the predictions of random graph models. Using a new algorithm for counting the number of shortest paths between vertices on a graph that pass through each other vertex, we have calculated the so-called betweenness measure of centrality

[8] On being informed of this latest honor, Prof. Rees is reported as replying, "I'm certainly relieved not to be the most disconnected astrophysicist" (H. Muir, *New Scientist,* November 25, 2000, p. 10).

on our graphs. We have also shown that for most authors the bulk of the paths between them and other scientists in the network go through just one or two or their collaborators, an effect that Strogatz has dubbed "funneling."

We have suggested a measure of the closeness of collaborative ties that takes account of the number of papers a given pair of scientists have written together, as well as the number of other coauthors with whom they wrote them. Using this measure we have added weightings to our collaboration networks and used the resulting networks to find those scientists who have the shortest average distance to others. Generalization of the betweenness and funneling calculations to these weighted networks is also straightforward.

The calculations presented in this article inevitably represent only a small part of the investigations that could be conducted using large network datasets such as these. We hope, given the high current level of interest in network phenomena, that others will find many further uses for collaboration network data.

Acknowledgments

The author would particularly like to thank Paul Ginsparg for his invaluable help in obtaining the data used for this study. The data used were generously made available by Oleg Khovayko, David Lipman, and Grigoriy Starchenko (Medline), Paul Ginsparg and Geoffrey West (Physics E-print Archive), Heath O'Connell (SPIRES), and Carl Lagoze (NCSTRL). The Physics E-print Archive and NCSTRL are maintained by Cornell University, while Medline and SPIRES are maintained by the National Center for Biotechnology Information and the Stanford Linear Accelerator Center, respectively.

In addition, the author would like to thank Steve Strogatz for suggesting the funneling effect calculation of Sec. 4.2, and Dave Alderson, László Barabási, Sankar Das Sarma, Paul Ginsparg, Rick Grannis, Jon Kleinberg, Laura Landweber, Sidney Redner, Ronald Rousseau, Steve Strogatz, Duncan Watts, and Douglas White for many useful comments and suggestions. This work was funded in part by the National Science Foundation under grant number DMS–0234188, by Intel Corporation, and by the Santa Fe Institute.

References

Abello, J., Buchsbaum, A., and Westbrook, J., 1998. A functional approach to external graph algorithms. In *Proceedings of the 6th European Symposium on Algorithms*. Springer, Berlin.

Adamic, L. A., Lukose, R. M., Puniyani, A. R., and Huberman, B. A., 2001. Search in power-law networks. *Phys. Rev. E* **64**, 046135.

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B., 1993. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Upper Saddle River, New Jersey.

Albert, R., Jeong, H., and Barabási, A.-L., 1999. Diameter of the world-wide web. *Nature* **401**, 130–131.

Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E., 2000. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152.

Barabási, A.-L., 2002. *Linked: The New Science of Networks*. Perseus, Cambridge, MA.

Barabási, A.-L., Jeong, H., Ravasz, E., Néda, Z., Schuberts, A., and Vicsek, T., 2002. Evolution of the social network of scientific collaborations. *Physica A* **311**, 590–614.

Batagelj, V. and Mrvar, A., 2000. Some analyses of Erdős collaboration graph. *Social Networks* **22**, 173–186.

Bernard, H. R., Killworth, P. D., Evans, M. J., McCarty, C., and Shelley, G. A., 1988. Studying social relations cross-culturally. *Ethnology* **2**, 155–179.

Bollobás, B., 1980. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics* **1**, 311–316.

Bollobás, B., 2001. *Random Graphs*. Academic Press, New York, 2nd ed.

Bordens, M. and Gómez, I., 2000. Collaboration networks in science. In H. B. Atkins and B. Cronin (eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Information Today, Medford, NJ.

Brandes, U., 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* **25**, 163–177.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J., 2000. Graph structure in the web. *Computer Networks* **33**, 309–320.

Chen, Q., Chang, H., Govindan, R., Jamin, S., Shenker, S. J., and Willinger, W., 2002. The origin of power laws in Internet topologies revisited. In *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE Computer Society.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C., 2001. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 2nd ed.

Davis, A., Gardner, B. B., and Gardner, M. R., 1941. *Deep South*. University of Chicago Press, Chicago.

Davis, G. F. and Greve, H. R., 1997. Corporate elite networks and governance changes in the 1980s. *Am. J. Sociol.* **103**, 1–37.

Ding, Y., Foo, S., and Chowdhury, G., 1999. A bibliometric analysis of collaboration in the field of information retrieval. *Intl. Inform. and Libr. Rev.* **30**, 367–376.

Ebel, H., Mielsch, L.-I., and Bornholdt, S., 2002. Scale-free topology of e-mail networks. *Phys. Rev. E* **66**, 035103.

Egghe, L. and Rousseau, R., 1990. *Introduction to Informetrics*. Elsevier, Amsterdam.

Erdős, P. and Kac, M., 1940. The Gaussian law of errors in the theory of additive number theoretic functions. *Am. J. Math.* **26**, 738–742.

Erdős, P. and Rényi, A., 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61.

Faloutsos, M., Faloutsos, P., and Faloutsos, C., 1999. On power-law relationships of the internet topology. *Computer Communications Review* **29**, 251–262.

Fararo, T. J. and Sunshine, M., 1964. *A Study of a Biased Friendship Network*. Syracuse University Press, Syracuse.

Freeman, L., 1977. A set of measures of centrality based upon betweenness. *Sociometry* **40**, 35–41.

Galaskiewicz, J. and Marsden, P. V., 1978. Interorganizational resource networks: Formal patterns of overlap. *Social Science Research* **7**, 89–107.

Goh, K.-I., Kahng, B., and Kim, D., 2001. Universal behavior of load distribution in scale-

free networks. *Phys. Rev. Lett.* **87**, 278701.

Grossman, J. W., 2002. The evolution of the mathematical research collaboration graph. *Congressus Numerantium* **158**, 202–212.

Grossman, J. W. and Ion, P. D. F., 1995. On a portion of the well-known collaboration graph. *Congressus Numerantium* **108**, 129–131.

Hoffman, P., 1998. *The Man Who Loved Only Numbers*. Hyperion, New York.

Kautz, H., Selman, B., and Shah, M., 1997. ReferralWeb: Combining social networks and collaborative filtering. *Comm. ACM* **40**, 63–65.

Kleinberg, J. M., 2000. Navigation in a small world. *Nature* **406**, 845.

Krapivsky, P. L. and Redner, S., 2001. Organization of growing random networks. *Phys. Rev. E* **63**, 066123.

Kretschmer, H., 1994. Coauthorship networks of invisible college and institutionalized communities. *Scientometrics* **30**, 363–369.

Lotka, A. J., 1926. The frequency distribution of scientific production. *J. Wash. Acad. Sci.* **16**, 317–323.

Łuczak, T., 1992. Sparse random graphs with a given degree sequence. In A. M. Frieze and T. Łuczak (eds.), *Proceedings of the Symposium on Random Graphs, Poznań 1989*, pp. 165–182. John Wiley, New York.

Mariolis, P., 1975. Interlocking directorates and control of corporations: The theory of bank control. *Social Science Quarterly* **56**, 425–439.

Melin, G. and Persson, O., 1996. Studying research collaboration using co-authorships. *Scientometrics* **36**, 363–377.

Milgram, S., 1967. The small world problem. *Psychology Today* **2**, 60–67.

Molloy, M. and Reed, B., 1995. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6**, 161–179.

Molloy, M. and Reed, B., 1998. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing* **7**, 295–305.

Moody, J., 2001. Race, school integration, and friendship segregation in America. *Am. J. Sociol.* **107**, 679–716.

Newman, M. E. J., 2001a. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98**, 404–409.

Newman, M. E. J., 2001b. Scientific collaboration networks: I. Network construction and fundamental results. *Phys. Rev. E* **64**, 016131.

Newman, M. E. J., 2001c. Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 016132.

Newman, M. E. J., 2001d. Clustering and preferential attachment in growing networks. *Phys. Rev. E* **64**, 025102.

Newman, M. E. J., 2003. The structure and function of complex networks. *SIAM Review* **45**, 167–256.

Newman, M. E. J. and Ziff, R. M., 2000. Efficient Monte Carlo algorithm and high-precision results for percolation. *Phys. Rev. Lett.* **85**, 4104–4107.

Newman, M. E. J., Strogatz, S. H., and Watts, D. J., 2001. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118.

Newman, M. E. J., Forrest, S., and Balthrop, J., 2002a. Email networks and the spread of computer viruses. *Phys. Rev. E* **66**, 035101.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H., 2002b. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **99**, 2566–2572.

O'Connell, H. B., 2000. Physicists thriving with paperless publishing. Preprint physics/0007040.

Padgett, J. F. and Ansell, C. K., 1993. Robust action and the rise of the Medici, 1400–1434. *Am. J. Sociol.* **98**, 1259–1319.

Pao, M. L., 1986. An empirical examination of Lotka's law. *Journal of the American Society for Information Science* (January 1986), 26–33.

Persson, O. and Beckmann, M., 1995. Locating the network of interacting authors in scientific specialties. *Scientometrics* **33**, 351–366.

Pool, I. de S. and Kochen, M., 1978. Contacts and influence. *Social Networks* **1**, 1–48.

Price, D. J. de S., 1965. Networks of scientific papers. *Science* **149**, 510–515.

Rapoport, A. and Horvath, W. J., 1961. A study of a large sociogram. *Behavioral Science* **6**, 279–291.

Redner, S., 1998. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* **4**, 131–134.

Ripeanu, M., Foster, I., and Iamnitchi, A., 2002. Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing* **6**, 50–57.

Scott, J., 2000. *Social Network Analysis: A Handbook*. Sage Publications, London, 2nd ed.

Seglen, P. O., 1992. The skewness of science. *J. Amer. Soc. Inform. Sci.* **43**, 628–638.

Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P. A., Mukherjee, G., and Manna, S. S., 2002. Small-world properties of the Indian railway network. Preprint cond-mat/0208535.

Strogatz, S. H., 2001. Exploring complex networks. *Nature* **410**, 268–276.

Voos, H., 1974. Lotka and information science. *Journal of the American Society for Information Science* (July-August 1974), 270–272.

Wasserman, S. and Faust, K., 1994. *Social Network Analysis*. Cambridge University Press, Cambridge.

Watts, D. J., 2003. *Six Degrees: The Science of a Connected Age*. Norton, New York.

Watts, D. J. and Strogatz, S. H., 1998. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., and Haythornthwaite, C., 1996. Computer networks as social networks. *Annual Review of Sociology* **22**, 213–238.

Ziff, R. M., Uhlenbeck, G. E., and Kac, M., 1977. The ideal Bose-Einstein gas, revisited. *Phys. Rep.* **32**, 169–248.