from *SIAM News*, Volume 35, Number 9, November 2002

Patterns of Collaboration in Mathematical Research

Six degrees of separation? Not for mathematicians, it seems, based on the author's study of a "collaboration graph," with publishing mathematicians as the vertices and two vertices joined by an edge if the mathematicians have ever written a joint paper.

By Jerrold W. Grossman

Each year mathematicians publish more than 50,000 research papers. Since 1940, *Mathematical Reviews* (MR; available electronically on the World Wide Web as MathSciNet [8]) has catalogued most of them, and MR's current database contains more than one and a half million items, produced by more than a third of a million authors. By studying this wealth of data, we can discern some interesting patterns of publication, and in particular some interesting patterns of collaboration.

For simplicity, we call each authored item in the MR database a "paper," although some of them are monographs of various kinds. We ignore nonauthored items in the database, such as conference proceedings—the relevant papers in the proceedings have their own entries as authored items. In maintaining this database, and making it available to subscribers in print form and on the Internet, the MR editors and staff have taken pains to identify authors as people and not merely as name strings—strings of characters that the journal listed as an author's name. For example, Raymond L. Johnson, Roberto Johnson, and Russell A. Johnson all published under the name string "R. Johnson," but each of the papers by "R. Johnson" in the database is identified with exactly one of these three people. To the extent that MR has been successful in this endeavor, the data will accurately reflect the publication habits and the social network of actual individuals. (Some errors of this type remain, to be sure, but we do not think they substantially affect our results. Indeed, before they corrected the mistake in 1995, MR listed a paper by the physicist Paul Erdős as being by the mathematician Paul Erdős. Now, MR denotes these two individuals Paul Erdős^{2} and Paul Erdős^{1}, respectively, using a convention that has become increasingly necessary. See [11] for more details.)

The data used in this article cover approximately the period from 1940 to 1999, inclusive, and we have broken it down approximately by decade. (There is necessarily a lot of imprecision in dating, partly because the reviews in MR typically appear about a year after publication, which in turn is often more than a year after submission.) We thank the American Mathematical Society for providing access to this data, as well as Patrick Ion of *Mathematical Reviews* for helpful conversations.

The cumulative data are summarized in Table 1, whose integer entries represent thousands. The left-most column includes all the data, and the remaining columns truncate the data after one or more decades. Data are given for all authors, as well as just for authors who have collaborated.

The third row of Table 1 shows the average number of papers per author. Since mathematicians at all stages in their careers are included, it is hard to know exactly how to view the statistic that the mean number of papers is about 7. Of course the distribution has a long right tail, with a standard deviation of more than 15. Table 2 shows the fractions of mathematicians who have written various numbers of papers. It can be seen from this table that just slightly more than half of all publishing mathematicians have published more than one paper, that the median number of papers is just 2, and that more than two thirds of us have written fewer than five papers. At the other extreme, eight people have written more than 500 papers apiece, including the legendary Paul Erdős, with about 1500 papers. When tenure committees count publications, this kind of information might help to put things in context.

	thru 90s	thru 80s	thru 70s	thru 60s	thru 50s	thru 40s
Number of papers	1598	1010	572	278	109	30
Number of authors	337	225	137	68	29	10
Mean papers/author	6.87	6.05	5.30	4.89	4.33	3.41
S.D. papers/author	15.34	12.91	11.26	10.49	8.88	5.70
Mean authors/paper	1.45	1.35	1.27	1.20	1.14	1.10
S.D. authors/paper	1.63	1.50	1.40	1.31	1.28	0.3
1-author papers	66%	73%	78%	84%	88%	91%
2-author papers	26%	22%	18%	13%	11%	8%
3-author papers	7%	4%	3%	2%	1%	1%
> 3-author papers	1%	1%	1%	1%	0%	0%
Collaborating authors	253	153	82	34	11	3
Fraction of all authors	75%	68%	60%	49%	39%	28%
Mean collaborators/author	2.94	2.26	1.67	1.20	0.83	0.49
Mean collaborators/ collaborating author	3.92	3.33	2.79	2.42	2.14	1.74

Table 1. *Cumulative data, by decade; integer figures represent thousands.*

Table 3 summarizes the data decade by decade, giving a better view of how things have changed over the years. Row 2 shows the explosion in the number of practicing mathematicians during the period we consider, a compounded annual growth rate of 6% per year (compared with a rate of less than 2% for the population of the world during the same period). We infer from row 3 that in the 1940s the mean number of papers per mathematician per year was about 0.3, that this figure grew to about 0.4 in the 1960s and 1970s, and that it reached nearly 0.5 in the 1990s.

We turn now to the issue of collaboration in mathematical research (which may, indeed, partially explain this increase in productivity). Mathematics is at neither extreme among the academic disciplines. Laboratory scientists tend to write articles with many authors; everyone who contributes to the experiments gets a credit. Scholars in the humanities usually engage in solitary work. In mathematics we find a definite trend toward increasing collaboration.

As Table 3 shows, the average number of authors per paper has gone from only 1.10 in the 1940s to 1.63 in the 1990s. During the 1940s only 28% of all publishing mathematicians wrote joint papers, whereas 81% of those who published in the 1990s collaborated at least once during that decade. In the 1940s and 1950s, nearly 90% of all papers were solo works, with only 1–2% of the papers having three or more authors. If we look at just the last two years' worth of items in the database, we find that by the late 1990s, fewer than half of all papers had just one author, and the number of papers with three or more authors had grown to 16%.

Number of papers	Fraction of mathematicians
1	42.7%
2	14.6%
3	8.0%
4	5.3%
5	3.9%
6–10	10.0%
11–20	7.4%
21–50	6.0%
51–100	1.7%
101–200	0.4%
> 200	< 0.1%

Table 2. *Fractions of mathematicians with various numbers of papers.*

	90s only	80s only	70s only	60s only	50s only	40s only
Number of papers	587	439	294	168	80	30
Number of authors	192	144	97	51	24	10
Mean papers/author	4.97	4.43	4.03	4.05	3.84	3.41
S.D. papers/author	8.31	6.91	6.15	6.60	6.73	5.70
Mean authors/paper	1.63	1.45	1.33	1.23	1.16	1.10
S.D. authors/paper	1.82	1.63	1.48	1.35	1.26	0.36
1-author papers	54%	66%	73%	81%	87%	91%
2-author papers	33%	27%	22%	16%	11%	8%
3-author papers	10%	6%	4%	2%	2%	1%
> 3-author papers	3%	1%	1%	1%	0%	0%
Collaborating authors	155	104	62	27	9	3
Fraction of all authors	81%	72%	64%	52%	41%	28%
Mean collaborators/author	2.84	2.16	1.62	1.18	0.84	0.49
Mean collaborators/ collaborating author	3.51	2.99	2.55	2.25	2.08	1.74

Table 3. *Data for each decade; integer figures represent thousands.*

To really get at the social phenomenon of collaboration in mathematical research, we construct the so-called collaboration graph, which we denote by C. The vertices of C are the 337,454 mathematicians in our database, and two vertices are joined by an edge if the two mathematicians have published a joint paper, with or without other coauthors. This gives us 496,489 edges, so the average degree of a vertex in C (the average number of coauthors per mathematician) is about 3. There are 84,115 isolated vertices in C (25%), which we will ignore for the purposes of this analysis; after all, these are not collaborating mathematicians. That leaves 253,339 vertices with degree at least 1. Viewed this way, the average degree (number of coauthors for a mathematician who collaborates) is about 4.

We look first at the degrees of the vertices—the distribution of the numbers of coauthors mathematicians have. Much recent research by mathematicians, physicists, sociologists, and others on large real-world networks (such as collaboration networks among scientists or film actors, the Internet, power grids, telephone call graphs, or neural networks of simple animals) suggests that the degrees usually follow a power law: The number of vertices of degree x is proportional to $x^{-\beta}$, where β is usually around 3. (See, for example, [1, 2, 3, 9, 12, 13]. Indeed, this has become a very "hot" area of research, with articles in *Nature* and the *Proceedings of the National Academy of Sciences* and several books for a general audience. The deeper mathematical questions involve ways to capture the structure of such networks with random graph models, especially as they evolve over time; the classic Erdős–Rényi model [4] does not apply.)

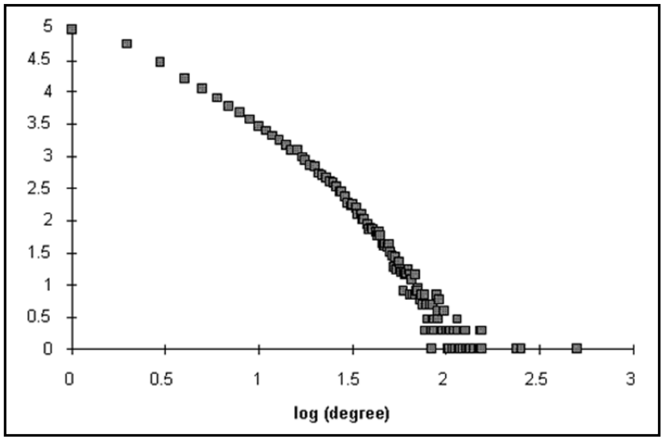

Figure 1. *Distribution of (nonzero) vertex degrees in* C.

Figure 1 shows a log–log plot of the frequencies of degrees, and the model seems to fit quite well. The data show that 37% of collaborating mathematicians have just one coauthor, 22% have two, 12% have three, 6% have four, and 23% have five or more. More than 400 mathematicians have written with more than 50 colleagues apiece, with Paul Erdős's 507 coauthors as the most extreme case. Again, the social interactions have increased over the years, no doubt due in part to electronic communication and the proliferation of conferences; Table 3 shows that the mean number of collaborators per mathematician in one decade grew from fewer than 1/2 in the 1940s to nearly 3 in the 1990s.

Other graphical properties of C also provide insight into the interconnectedness of mathematicians. For example, the collaboration graph has one giant component with 208,200 vertices and 461,643 edges; the remaining 45,139 nonisolated vertices and 34,846 edges split into 16,883 components, having from 2 to a maximum of 39 vertices (two thirds of the components are just isolated edges).

Next, we concentrate just on the giant component of C and consider the distribution of distances between vertices (number of edges in a shortest path joining the vertices). The average distance between two vertices is between 7 and 8, with a standard deviation of about 1½. Apparently, the appropriate popular buzz phrase for mathematicians should be "eight degrees of separation" [7].

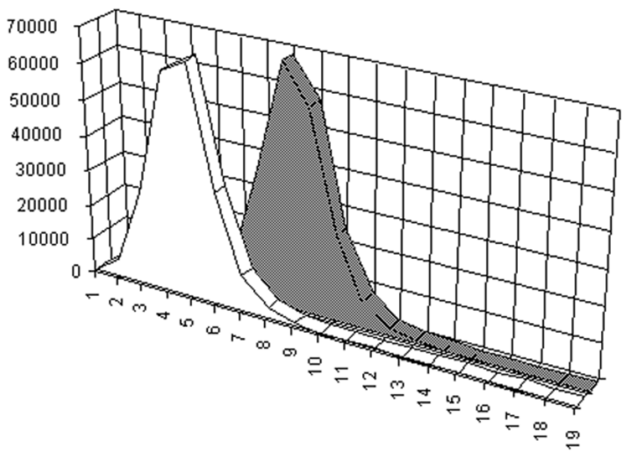

Figure 2. *Distributions of Erdős numbers (front) and Jane Doe numbers (back).*

The diameter of the giant component (maximum distance between two vertices) is 27, and the radius (minimum eccentricity of a vertex, with eccentricity defined as the maximum distance from that vertex to any other) is 14. For any fixed vertex u in the giant component, we can ask for the shape of the distribution of the distances from u to the other 208,199 vertices in this component. The distance from u to v is, of course, the familiar "Erdős number" of v when u = Erdős [5, 6]. These distributions are bell-shaped, usually with long right tails. The means typically range from 6 to 11 (although the mean is only 4.7 for Paul Erdős, and it goes as high as 17.5 for one person on the "fringes" of C). The standard deviations of these distributions are remarkably constant, with the numbers varying only between 1.19 and 1.35 in a random sample of 100 mathematicians. So although the *average* "Jane Doe" number varies quite a bit, depending on who Jane Doe is, the *distribution* of these numbers has pretty much the same shape and spread for everyone. Figure 2 shows the distribution of Erdős numbers and the distribution of Jane Doe numbers for a person chosen at random. It seems that people farther from the heart of the graph might take longer to get to the heart but, once there, have the same fan-out pattern.

As a final measure, we compute the clustering coefficient of C to be 0.15. The *clustering coefficient* [10] of a graph is the fraction of ordered triples of vertices a, b, c in which edges ab and bc are present that have edge ac present. In other words, how often are two neighbors of a vertex adjacent to each other? This value is 10,000 times higher than we would expect for a random graph with 253,000 vertices and 496,000 edges. Such behavior is typical of the "small-world" networks studied in the literature [12].

The *Mathematical Reviews* data provide a wonderful opportunity for further study of the publishing patterns of mathematicians, both as individuals and as a highly and intricately connected corpus. For instance, it would be interesting to look at the differences among mathematicians in different subfields, to see to what extent a person's publication record over the first six years gives an indication of future productivity, or to notice significant differences in publication or collabor-ation patterns among mathematicians at different types of institutions or in different countries. Further information is available on the author's Erdős Number Project Web site [5].

References

[1] W. Aiello, F. Chung, and L. Lu, *A random graph model for power law graphs*, Experiment. Math., 10 (2001), 53–66; MR 2001m:05233.

[2] A.-L. Barabási, *Linked: The New Science of Networks*, Perseus, New York, 2002.

[3] M. Buchanan, *Nexus: Small Worlds and the Groundbreaking Science of Networks*, W.W. Norton, New York, 2002.

[4] P. Erdős and A. Rényi, *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 5 (1960), 17–61; MR 23#A2338.

[5] J.W. Grossman, The Erdős Number Project, http://www.oakland.edu/~grossman/erdoshp.html, 2002.

[6] J.W. Grossman and P.D.F. Ion, *On a portion of the well-known collaboration graph*, Proceedings of the Twenty-sixth Southeastern International Conference on Combinatorics, Graph Theory and Computing, Boca Raton, Florida, 1995, Congressus Num-erantium, 108 (1995), 129–131; CMP 1 369 281.

[7] J. Guare, *Six Degrees of Separation*, Random House, New York, 1990.

[8] MathSciNet, *Mathematical Reviews* on the Web, 1940–present, American Mathematical Society, http://www.ams.org/mathscinet.

[9] M.E.J. Newman, *The structure of scientific collaboration networks*, Proceedings of the National Academy of Sciences USA, 98 (2001), 404–409; CMP 1 812 610.

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

[11] B. TePaske-King and N. Richert, *The identification of authors in the* Mathematical Reviews *database*, Issues in Science and Technology Librarianship, 31 (Summer 2001); http://www.library.ucsb.edu/istl/01-summer/databases.html.

[12] D.J. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton University Press, Princeton, New Jersey, 1999; MR 2001a:91064.

[13] D.J. Watts and S.H. Strogatz, *Collective dynamics of "small-world" networks*, Nature, 393 (1998), 440–442.

Jerrold W. Grossman is a professor in the Department of Mathematics and Statistics at Oakland University, in Rochester, Michigan.