Patterns of Research in Mathematics

Jerrold W. Grossman

For ten days in April 2004, an Erdős number was for sale on the World Wide Web auction site eBay [9]. To be more precise, a scientific consultant in Ann Arbor, Michigan, was selling 40 hours of his time to the winning bidder, with the goal of publishing a jointly authored research article. Because the seller has Erdős number 4, the buyer will obtain Erdős number 5 if the project is successful. Seventeen people participated in the auction, and the winning bid was over $1,000. (The seller is quite serious about this, as part of his crusade for encouraging collaboration in research. Some people find it harmless fun. A few are outraged; in particular, the high bidder posted a message, explaining that he won the auction "not because I intend to pay or to collaborate with the seller—my Erdos [sic] number is already 3—but to stop the mockery this person is doing of the paper/journal system.")

The notion of a researcher's Erdős number has circulated among mathematicians for decades [3, 11]. It is simply the length of (number of edges in) the shortest path from Paul Erdős (1913–1996) to the researcher in the collaboration graph—the graph whose vertices are researchers and whose edges join any two people who are coauthors on a published research article or book. Because Erdős had by far the most collaborators of any mathematician (509 at latest count), he is the natural "center" for the collaboration graph, at least in mathematics. The much-visited Erdős Number Project website [5] lists Erdős's coauthors, as well as the nearly 7,000 people with Erdős number 2, and provides much additional information.

Interest in Erdős numbers illustrates mathematicians' curiosity about how we go about doing what we do, and in particular about the social aspects of our profession. The main purpose of this article is to make better known some facts and figures about the world of mathematical research. The interested reader should also consult a recent article in *SIAM News* [6], which complements much of what is discussed here, as well as the website referenced above. The data we used were kindly provided by the American Mathematical Society and cover approximately the time period 1940–1999.

How Much Research Is Going On?

The Society's *Mathematical Reviews* (MR) currently catalogs (and in most cases publishes reviews or edited author summaries of) about 86,000 published items per year that can generally be classified as research in the mathematical sciences. At the turn of the century, the database contained about 1.6 million papers (and books), produced by about 300,000 authors. (Two notes: 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 [8], the MR staff has taken pains to identify authors as people and not merely as name

Jerry Grossman is professor of mathematics at Oakland University, Rochester, MI. His email address is grossman@oakland.edu.

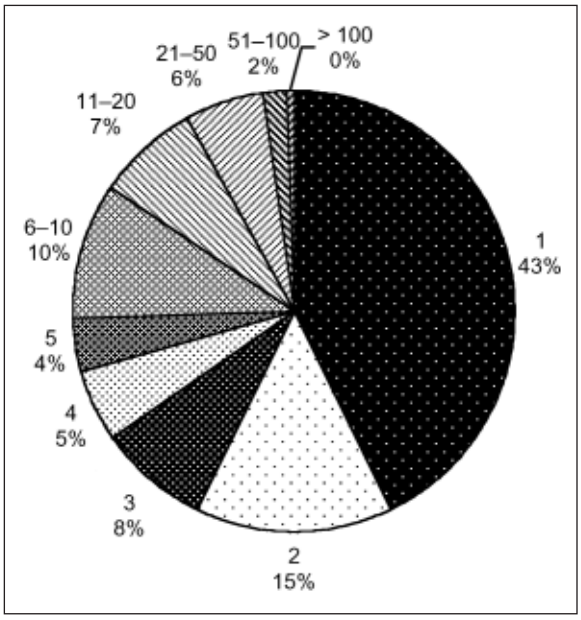


Figure 1. The distribution of the total number of papers per author.

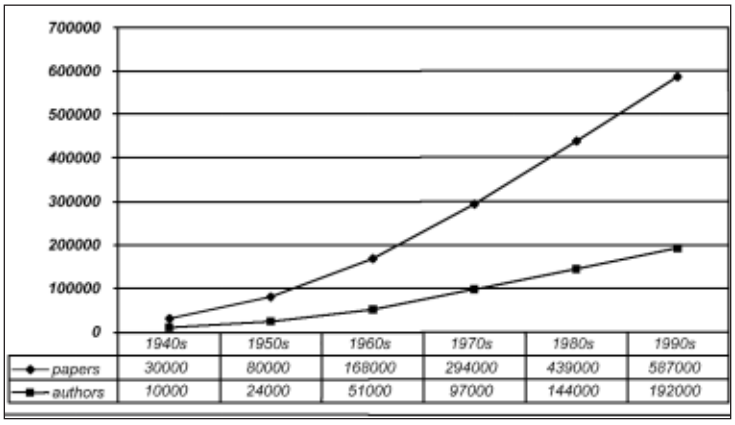


Figure 2. The increasing quantity of mathematics and mathematicians.

strings—see [12] for details. Although some misidentifications are present in the database, we do not think they substantially affect our results.)

Figure 1 shows the number of authors in the database with different numbers of papers. About 43% of all authors have just one paper. The median is 2, the mean 6.87, and the standard deviation 15.35. It is interesting (for tenure review committees?) to note that the 60th percentile is 3 papers, the 70th percentile is 4, the 80th percentile is 8, the 90th percentile is 17, and the 95th percentile is 30. Because the database is dynamic, these numbers must be viewed with caution; someone who has published only one paper as of today may have many more coming down the road.

The author list contains over 1,500 people with more than 100 papers, including eight mathematicians with more than 500 papers: Paul Erdős with 1,401, Drumi Bainov with 782, Leonard

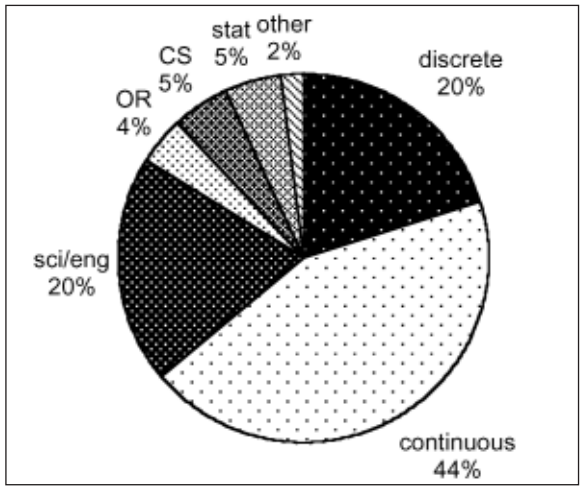


Figure 3. The distribution of papers by groups of MSC categories.

Carlitz (1907–1999) with 730, Lucien Godeaux (1887–1975) with 644, Saharon Shelah with 600, Hari M. Srivastava with 537, Frank Harary with 534, and Richard Bellman with 522. (Many of these counts have increased in the past five years.)

Not surprisingly, research productivity has increased over time. Figure 2 shows the number of papers and number of authors in each of the decades covered by our data. The mean number of papers per author (within the given decade) increased gradually, from 3.4 in the 1940s to 5.0 in the 1990s.

Is There a Difference in Output among Areas of Mathematics?

Beginning in 1980, it became easier to identify the primary "area" of mathematics for each paper, because the MR number for a paper encoded the primary subject classification. In order to draw some summary conclusions, we divide the mathematics subject classification (MSC) categories into groups, admittedly somewhat arbitrary, as shown in Table 1. (A few adjustments had to be made to account for changes in the classification scheme over time. For example, the old Section 10 was replaced by Section 11 in 1984, and papers in the former are included in our counts for the latter. New categories added in 2000 are not included.) The data for this section are the 886,000 papers and 220,000 authors from 1980 to 1999.

Figure 3 shows the fraction of papers in each group. It probably comes as no surprise to anyone hanging around major mathematics departments that continuous mathematics has a clear plurality. Of course there are thousands of papers in the science and engineering category that fall outside the scope of MR and are therefore not included.

At a finer level of detail, we see in Table 2 the particular sections that include more than 3% of the total. These 11 sections together account for 46% of the papers in MR.

Many mathematicians work in more than one area. Figure 4 shows the fraction of authors, among the 130,000 who have published more than one paper during 1980–1999, who published in one or more areas during that time. (The 90,000 authors having just one paper are omitted from this chart.) Breadth is a two-edged sword, and it is certainly not to be assumed that the more areas one works in, the better.

One might wonder whether researchers in some areas of mathematics publish more papers than researchers in other areas. (Some might claim that this would mean that such researchers are better at finding new theorems. Skeptics might argue that the cultural climate surrounding some areas leads to the acceptance of more trivial results.) Figure 5 shows the mean number of papers per author in each group. These figures are lower than the overall mean of 6.87 for two possible reasons: The data cover only the time period 1980–1999, and a researcher who publishes in more than one group is counted here in each group separately.

There does not seem to be a large difference in output between continuous mathematics and discrete mathematics. Furthermore, the individual sections with the most papers per author (between 4.25 and 4.62) seem to range randomly over the groupings, and include, in decreasing order, 05, 16, 60, 14, 81, 83, 35, 54, 53, and 20. Given that many researchers in peripheral areas, such as statistics or engineering, may have many papers not included in MR, it does not seem to be significant that the paper counts for these groups are lower.

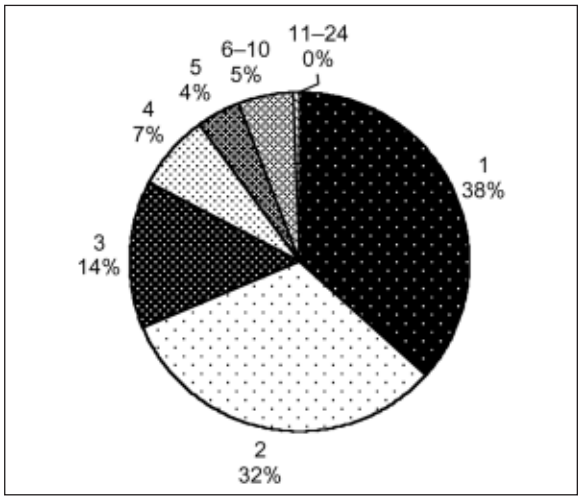


Figure 4. The distribution of the number of areas in which authors with more than one paper published.

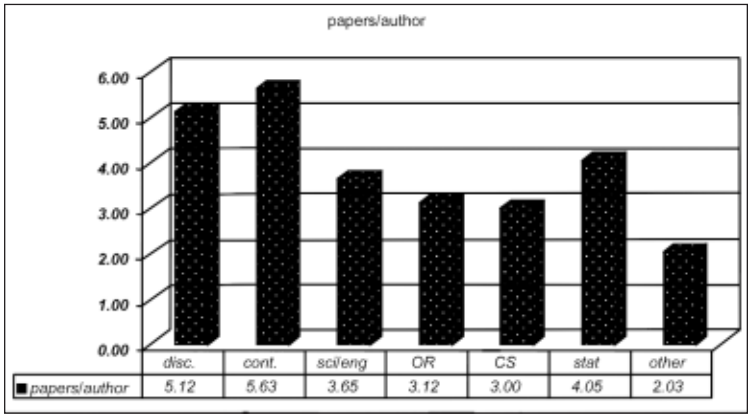


Figure 5. The mean number of papers per author, by group.

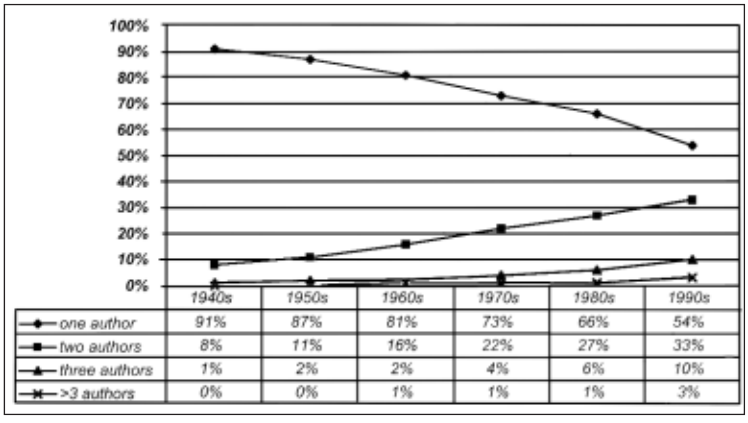


Figure 6. The fraction of papers with a given number of authors, by decade.

How Much Collaboration Is Going On, and Does Area Matter?

About half of the papers currently being published have more than one author. This has not always been the case. Figure 6 shows the fractions of papers in each decade with one, two, three, or more than three authors. For the database as a whole, 66% of the papers have one author, 26% have two, 7% have three, and 1% have four or more. The mean number of authors per paper has risen from 1.10 in the 1940s to 1.63 in the 1990s. The mean for the entire database is 1.45. About 75% of all authors appearing in the database have written joint papers.

To determine whether area of mathematics correlates with multiauthorship, we look again at the 1980–1999 data broken down by the MR sections and the groupings shown in Table 1. During this time period, the mean number of authors per paper was about 1.52, and about 39% of the papers were joint work. Figure 7 shows the extent of collaboration by group. The mean number of authors per paper is very close in the pure mathematics areas (1.45 for continuous, 1.41 for discrete). As might have been guessed, this parameter is higher (around 1.7) for MR papers in computer science, science, and engineering; it is significantly lower for the "Other" group.

Table 3 shows the individual MR sections with the most and the least collaboration. Apparently it is not fashionable to have too many authors on a paper in certain areas of mathematics.

Pure "discrete" mathematics

03 Logic, set theory, foundations
05 Combinatorics
06 Order, lattices, ordered alg. struct.
08 General algebraic systems
11 Number theory
12 Field theory and polynomials
13 Commutative rings and algebras
14 Algebraic geometry
15 Linear, multilinear alg.; matrix theory
16 Associative rings and algebras
17 Nonassociative rings and algebras
18 Category theory; homological algebra
19 *K*-theory
20 Group theory and generalizations
51 Geometry
52 Convex and discrete geometry

Science and engineering

70 Mechanics of particles and systems
73 Mechanics of solids
76 Fluid mechanics
78 Optics, electromagnetic theory
80 Class. thermodynamics, heat transfer
81 Quantum theory
82 Stat. mechanics, structure of matter
83 Relativity and gravitational theory
85 Astronomy and astrophysics
86 Geophysics
92 Biology and other natural sciences
93 Systems theory; control

Other

00 General
01 History and biography

Pure "continuous" mathematics

22 Topological groups, Lie groups
26 Real functions
28 Measure and integration
30 Functions of a complex variable
31 Potential theory
32 Sev. complex vars., analytic spaces
33 Special functions
34 Ordinary differential equations
35 Partial differential equations
39 Difference and functional equations
40 Sequences, series, summability
41 Approximations and expansions
42 Fourier analysis
43 Abstract harmonic analysis
44 Integral transforms, oper. calc.
45 Integral equations
46 Functional analysis
47 Operator theory
49 Calc. of variations, optimal control
53 Differential geometry
54 General topology
55 Algebraic topology
57 Manifolds and cell complexes
58 Global analysis, analysis on manifolds
60 Probability theory, stochastic proc.
65 Numerical analysis

Operations research

90 Oper. res., math. programming

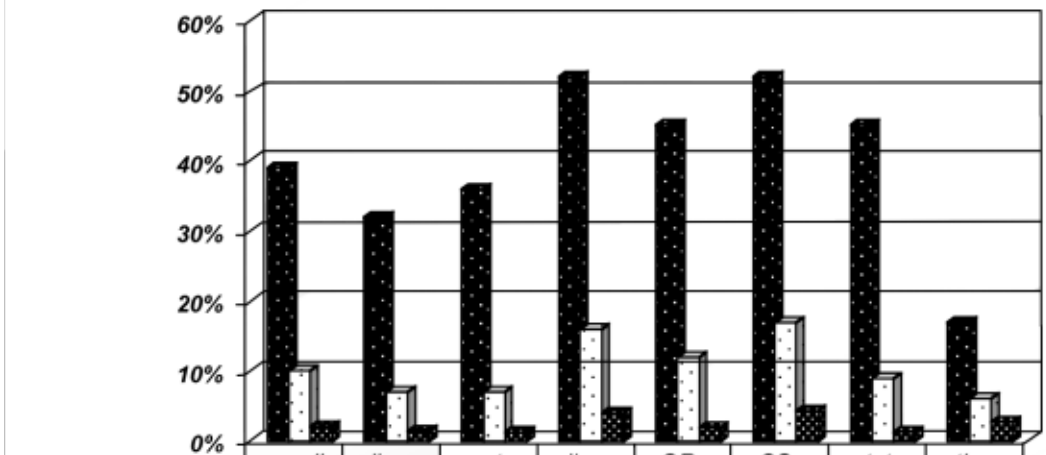
Statistics

62 Statistics

Computer science

68 Computer science
94 Information, circuits, communication

Table 1. Groups of subjects in the Mathematics Subject Classification.



	overall	disc.	cont.	sci/eng	OR	CS	stat	other
joint papers	39%	32%	36%	52%	45%	52%	45%	17%
>2 authors	10%	7%	7%	16%	12%	17%	9%	6%
>3 authors	2.0%	1.4%	1.2%	3.9%	1.9%	4.3%	1.3%	2.7%

Figure 7. The extent of collaboration by group, 1980–1999.

81	Quantum theory	5.2%
35	Partial differential equations	5.2%
62	Statistics	4.6%
65	Numerical analysis	4.6%
60	Probability theory, stochastic proc.	4.3%
90	Oper. res., math. programming	4.3%
58	Global analysis, analysis on manifolds	4.2%
05	Combinatorics	3.6%
68	Computer science	3.6%
34	Ordinary differential equations	3.2%
11	Number theory	3.2%

Table 2. MSC categories with the most papers.

What Does the Collaboration Graph Look Like?

The past several years have seen an explosion in the interest in properties of large real-world networks. Graph theorists are not the only leading players here; physicists, computer scientists, and sociologists are turning out papers by the dozens. (See, for example, an excellent comprehensive survey by physicist Mark Newman [10].) These graphs—whether social networks such as collaboration graphs or telephone call graphs, information networks such as citation graphs or the links on the World Wide Web, technological networks such as the Internet or power grids, or biological networks such as protein interactions or the neural network of a worm—do not display the kinds of properties that "almost all" graphs do. Instead, parameters such as the degree distribution (how many edges are incident to each vertex), the average distance between vertices, and the degree of clustering of edges (as opposed to their being randomly distributed throughout the graph) exhibit interesting patterns. (Fittingly, Paul Erdős led the early research in the study of random graphs [4].)

Here are a few facts about C, the mathematics research collaboration graph for 1940–1999, according to MR data. (Of course this is just a subgraph of the collaboration graph for all disciplines; see the final section of this article.) The graph C has about 337,000 vertices and 496,000 edges, so the average number of collaborators per person is 2.94. The graph has one large connected component consisting of about 208,000 vertices. Of the remaining 129,000 authors, 84,000 of them have written no joint papers (these are isolated vertices in C). The other 45,000 are distributed in 17,000 components with up to 39 vertices, but 11,000 of these components have just two authors. The average number of collaborators for people who have collaborated is 3.92; the average number of collaborators for people in the large component is 4.43; and the average number of collaborators for people who have collaborated but are not in the large component is 1.54.

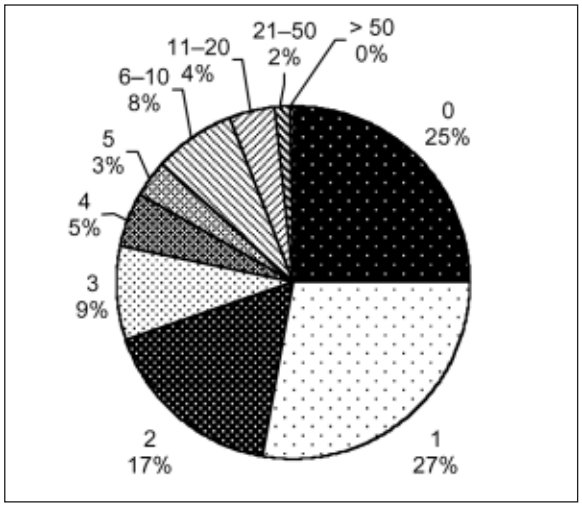


Figure 8. The distribution of the number of collaborators each mathematician has.

Figure 8 shows the distribution of the number of collaborators per mathematician. In graph-theoretical terms, this chart tabulates the degrees of the vertices in C. The median is 1, the mean is 2.94, and the standard deviation is 5.50. (If we omit the isolated vertices, then the median degree is 2, the mean is 3.92, and the standard deviation is 6.04.) Recent research (see [10]) has indicated that we should expect the nonzero degrees to follow a power law: The number of vertices with degree d should be roughly proportional to $d^{-\beta}$, where β is somewhere around 3. Indeed, when we fit such a model to our data (grouping the data in the tail), we find the exponent to be about 2.97, with a correlation coefficient for the model of $r = 0.97$. A slightly more accurate model throws in an exponential decay factor, and with this factor present, the exponent is 2.46, and $r = 0.98$. Apparently these models are appropriate for these data.

Forty-four people in the database have more than 100 collaborators, led by Paul Erdős with 502, Frank Harary with 254, and Yuri Alekseevich Mitropolskii with 240. (Harary is an Erdős coauthor, and Mitropolskii's Erdős number is 3.) The next six most sociable mathematicians (all of whom have between 152 and 156 coauthors listed in MR through 1999), with their Erdős numbers shown in parentheses, are Noga Alon (1), Andrei Nikolaevich Kolmogorov (1903–1987) (4), Saharon Shelah (1), Sergei Petrovich Novikov (3), Aleksandr Andreevich Samarskii (3), and Hari M. Srivastava (2). (Again, most of these counts have grown in the past five years.)

The distance $d(u, v)$ between two vertices in a graph is just the number of edges in a shortest path from one vertex to the other. The eccentricity of vertex u is $e(u) = \max_v d(u, v)$, the diameter of the graph is $\max_{u,v} d(u, v)$, and the radius of the graph is $\min_u e(u)$. The large component of C has radius 14 and diameter 27. There are at least three

MR Section		mean au.	>1 au.	>2 au.	>3 au.
68	Computer science	1.77	53%	17.7%	4.7%
	(all science and engineering sections)	1.73	52%	16.4%	3.9%
94	Information, circuits, communication	1.67	50%	13.5%	3.2%
05	Combinatorics	1.64	46%	13.7%	3.2%
65	Numerical analysis	1.61	46%	12.3%	2.2%
90	Oper. res., math. programming	1.59	45%	11.6%	1.9%
33	Special functions	1.58	45%	9.6%	2.0%
62	Statistics	1.56	45%	8.7%	1.3%
58	Global analysis, analysis on manifolds	1.55	40%	11.3%	2.7%
39	Difference and functional equations	1.52	41%	9.1%	1.4%
	(average over all sections)	**1.52**	**39%**	**9.9%**	**2.0%**
51	Geometry	1.34	28%	4.9%	0.7%
19	K-theory	1.33	26%	5.7%	1.3%
18	Category theory; homological algebra	1.33	28%	4.5%	0.4%
11	Number theory	1.32	26%	5.0%	0.7%
31	Potential theory	1.32	27%	5.0%	0.3%
14	Algebraic geometry	1.31	26%	4.4%	0.6%
03	Logic, set theory, foundations	1.30	24%	5.1%	0.9%
32	Sev. complex vars., analytic spaces	1.30	26%	3.7%	0.5%
12	Field theory and polynomials	1.30	25%	4.2%	0.5%

Table 3. MSC categories with the most and least collaboration.

vertices with eccentricity 14 (including Noga Alon, but not including Paul Erdős, whose eccentricity is 15). By randomly sampling 66 pairs of vertices, we found the average distance between two vertices to be around 7.8, with a standard deviation of 1.4. The median of the sample was 8, with the quartiles at 6.75 and 9. The smallest and largest distances in the sample were 5 and 12, respectively. The appropriate phrase for C, then, is perhaps "nine degrees of separation" [7], if we wish to account for three quarters of all pairs of mathematicians.

The clustering coefficient of a graph is defined as the fraction of ordered triples of vertices u, v, w in which edges uv and vw are present that have edge uw present. (In other words, how often are two neighbors of a vertex adjacent to each other?) The clustering coefficient of the collaboration graph is $913659/6072790 = 0.15$. The fairly high value of this parameter (it would be about 10^{-5} if the edges occurred at random), together with the fact that path lengths are small, indicates that C is a "small world" graph [13].

The discussion above refers to what might be called the collaboration graph of the first kind, in which two authors are joined by an edge if they have published a joint paper, regardless of how many other coauthors that paper has. A graph C' with fewer edges could be constructed by purists who put an edge between u and v only if u and v have published a two-author paper together. See [5] for comparable information about this collaboration graph of the second kind. An intriguing subgraph of C' is its main core. The k-core of a graph is the (unique) largest subgraph all of whose vertices have degree at least k; the smallest nonempty k-core (i.e., the one for largest k) is its main core [1]. For C' the main core is the 5-core, and it has 44 vertices and 162 edges: Paul Erdős, 36 Erdős coauthors, and seven people with Erdős number 2. Furthermore, because the core contains several copies of K_5 (the complete graph on five vertices) and $K_{3,3}$ (the complete bipartite graph with three vertices in each part), we know that C' is nonplanar (thereby answering a question raised by Erdős).

And What Is Your Erdős Number?

Mathematicians like to have fun computing their Erdős numbers. (Again, we could consider Erdős numbers of the first or second kinds. Here we will restrict ourselves to the former; see [5] for further information on the latter, as well as an updating of many of the statistics mentioned in this article using data through 2004.) Their distribution (for the 208,000 mathematicians in the large component of C in the MR database) is shown in Figure 9. The median Erdős number is 5 (as is the mode), the mean is 4.69, and the standard deviation is 1.27.

The Erdős numbers of the 44 Fields medalists are all 5 or less, with the exception of Laurent Lafforgue, whose publication list in MR shows no coauthors at all. Comparable statements can be made for winners of other honors, such as the Wolf Prize in Mathematics. A few mathematicians from the nineteenth and early twentieth centuries formally collaborated, and we have been able to establish finite Erdős numbers for such people as David Hilbert (4) and Ferdinand Georg Frobenius (3), but not, unfortunately, for Bertrand Russell. The

author would be interested in learning of other results of this type.

One could certainly extend the collaboration graph outside the field of mathematics. The eminent biologist Eugene V. Koonin, at the National Center for Biotechnology Information, has Erdős number 2 (via a paper with László Székely and others on genomes), and this leads to small finite Erdős numbers for many scientists. Indeed, it is probably possible to connect to Erdős a large fraction of people who have published in the biological sciences. With a couple of hours' work on the Web, the author was able to establish an upper bound of 9 for the Erdős number of his brother, a practicing physician, who was a coauthor on a single biology paper resulting from a summer internship. Similar results should hold for physics, chemistry, and computer science. It would be interesting to explore how the issues raised in this article apply to research in the social sciences and the humanities.

Microsoft founder Bill Gates has Erdős number 4, as does Linus Pauling, who won the 1954 Nobel Prize in Chemistry and the 1962 Nobel Peace Prize. It is not hard to find small finite Erdős numbers for many other Nobel prize winners as well, from Albert Einstein (2; 1921 Physics) to Francis H. C. Crick (7; 1962 Medicine) to Herbert A. Simon (3; 1978 Economics). Further trivia of this sort can be found on the Erdős Number Project website, as well as in [2].

The author has implemented a breadth-first search in the collaboration graph and would be happy to assist curious readers in determining their own Erdős numbers. Furthermore, MathSciNet has recently added a feature that allows users to find the distance in C between any two authors.

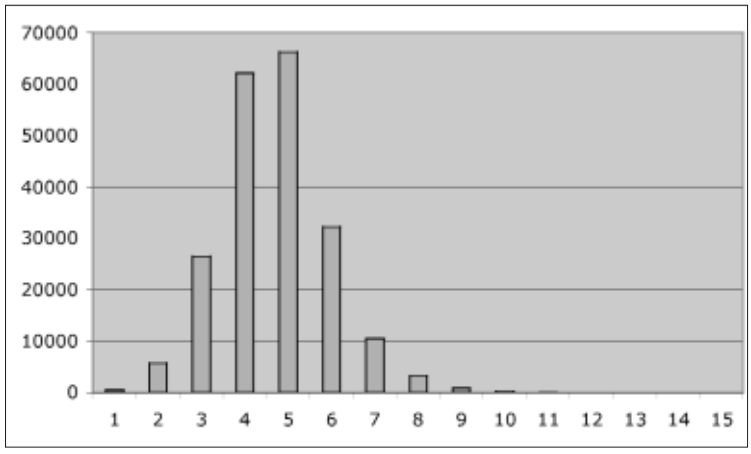


Figure 9. The number of people with different Erdős numbers.

References

[1] Vladimir Batagelj and Andrej Mrvar, Some analyses of Erdős collaboration graphs, *Social Networks* **22** (2000), 173–186; **MR** 2001j:91101.

[2] Rodrigo De Castro and Jerrold W. Grossman, Famous trails to Paul Erdős, with a sidebar by Paul M. B. Vitanyi, *Math. Intelligencer* **21:3** (1999), 51–63; **MR** 2000j:01053.

[3] Casper Goffman, And what is your Erdős number?, *American Mathematical Monthly* **76** (1969), 791.

[4] P. Erdős and A. Rényi, On random graphs, I, *Publ. Math. Debrecen* **6** (1959), 290–297; **MR** 22#10924.

[5] Jerrold W. Grossman, The Erdős Number Project, `http://www.oakland.edu/enp`, 1996–present.

[6] ____ , Patterns of collaboration in mathematical research, *SIAM News* **35:9** (November, 2002), 1 and 8–9.

[7] John 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] Richard Monastersky, Hot type: Co-author for sale, *The Chronicle of Higher Education* **50:38** (May 28, 2004), A15.

[10] M. E. J. Newman, The structure and function of complex networks, *SIAM Review* **45** (2003), 167–256; **MR** 2010377.

[11] Tom Odda [pseudonym for Ronald L. Graham], On properties of a well-known graph or what is your Ramsey number?, *Topics in graph theory* (New York, 1977), 166–172, Ann. New York Acad. Sci. **328**, New York Acad. Sci., New York, 1979; **MR** 81d:05055.

[12] Bert TePaske-King and Norman 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`.

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