The Evolution of the Mathematical Research Collaboration Graph

Jerrold W. Grossman
Department of Mathematics and Statistics
Oakland University
Rochester, MI 48309-4485
e-mail: grossman@oakland.edu

Abstract

We discuss some properties of the research collaboration graph for mathematicians, look at its evolution over time, and survey some random models that might produce graphs of this sort. Our approach is more experimental and statistical than theoretical. Further information is available on the Erdős Number Project web site (especially the subpage http://www.oakland.edu/~grossman/trivia.html).

AMS Subject Classification (2000):
Primary: 91D30
Secondary: 01A80, 05C80, 05C90, 94C15

§1. Introduction

The mathematics research collaboration graph C_{math} has as its vertices all mathematicians who have published research papers. Two vertices are joined by an edge in C_{math} if the two mathematicians have published a joint paper, with or without other coauthors. Using 60 years of data from *Mathematical Reviews* (MR, available electronically on the World Wide Web as MathSciNet [15]), from its beginning in 1940, we find that this graph currently has about 337,000 vertices and 496,000 edges.

Traditional models of random graphs do not produce graphs that look at all like C_{math}. For example, in C_{math} (and similar "small world" graphs in the sense of Watts [23]) the number of vertices of a given degree is approximately proportional to a power (somewhere around -3) of the degree, whereas the traditional model gives a Poisson distribution. Recently a number of researchers in both mathematics and physics have suggested various random graph models to explain the structure and evolution of graphs like C_{math}.

In this paper we look at some details of the structure of C_{math}. In addition to discussing the entire graph as it exists at the present time, we also study its evolution over the past 60 years. We believe that these data are interesting in their own right as a reflection of the way mathematical research gets done, apart from the mathematical questions raised about how to model and analyze social interactions. We also observe to what extent the structure of C_{math} conforms to the predictions made by the various random graph models that have been proposed.

§2. The MR Data and the Construction of C_{math}

The American Mathematical Society's *Mathematical Reviews* database consists of about 1.6 million authored items (mainly research papers published in peer-refereed journals), written by a total of about 337,000 different authors. (These are the figures as of the end of the 20th century; of course these numbers continue to grow.) For simplicity we call each such item a "paper". (We ignore non-authored 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 collaboration graph will accurately reflect the social network of individuals and not accidents caused by misidentification. (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, these two individuals are identified as Paul Erdős^2 and Paul Erdős^1, respectively, using a convention that has become increasingly necessary. See [21] for more details.)

This database of authored items gives rise to a bipartite graph B_{math}, whose vertices of one type are the papers and vertices of the other type are the authors, with an edge between a paper and each of its authors. The graph B_{math} has about 2.3 million edges, from which it follows that the average number of authors per paper is about $1\frac{1}{2}$, and the average number of papers per author is about 7. (The latter distribution is heavily skewed, with first quartile 1, median 2, third quartile 6, standard deviation about 15, and maximum 1401, the number of papers by Paul Erdős in the database.) Any bipartite graph gives rise to two association graphs by squaring and restricting the vertices to being of one type or the other. Thus the collaboration graph C_{math} has the set of authors as its vertices, with two authors adjacent if they are among the authors on some paper—in other words, if they are published research collaborators. (We have also studied the graph obtained when we restrict the set of papers to be only those with exactly two authors; see [11] for more discussion of this "collaboration graph of the second kind".)

We corrected a few anomalies in C_{math} by hand before analyzing it. For example, we removed the author that MR identified as "et al.", who was on the author list of a number of papers, including one with no coauthors! Based on extensive experience with the database over the past seven years, we are confident that problems of this sort do not significantly distort the true picture of the collaboration graph.

§3. Models for Small World Graphs

The mathematics collaboration graph (as well as other social networks studied in the

literature) exhibits several interesting features. First, although the number of edges is fairly small (just a little larger than the number of vertices), the average path length between vertices in the same component is small. Furthermore, there is a "giant component" of the graph that encompasses a majority of the authors, and the remaining components are very tiny. Second, the degrees of the vertices in the collaboration graph follow a "power law" pattern—the number of vertices of degree x is proportional to a (negative) power of x. Third, the clustering coefficient is fairly large. (The *clustering coefficient* [20] 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?) The question, then, is, "What model of random graphical evolution will produce graphs with these (and other) properties of the collaboration graph?"

The first model for constructing random graphs [10] appeared in 1961, when Erdős and Rényi fixed the (large) number of vertices (n) and the number of edges (m, which could depend on n) and chose endpoints of edges uniformly at random until the required number of edges had been obtained. Beautiful and surprising theorems describe how, almost surely, the structure of the resulting graph depends only on how large m is, relative to n. For example, if $m > n/2$, then there will be a giant component containing most of the vertices, and if $m > (n \log n)/2$, then the graph will be connected.

The Erdős-Rényi model is not suitable for describing collaboration graphs, however, because collaborations are not formed uniformly at random. For example, if u and v have each collaborated with the same person, then it makes sense that u and v are more likely to have collaborated with each other than if they do not share a common collaborator. (This will tend to make the clustering coefficient higher.) Furthermore, it seems likely that people who have already collaborated with many people are more likely to collaborate with someone else than those who have few collaborators. (Once vertices attain a high degree, their degree will tend to increase even more as the graph evolves, so the distribution of degrees will be skewed.) Thus in a realistic model, the probability of adding edge uv in the construction process should depend on such things as how close u and v already are in the graph constructed so far, or how many edges are already incident to each of them. Different models are needed.

Both mathematicians and physicists have proposed a variety of interesting models. One way to guarantee the power law pattern of vertex degrees is to fix the degree sequence according to such a law and then construct the graph to have exactly, or expectedly, the given degree sequence. Specifying degrees ahead of time was first tried in 1978 by Bender and Canfield [6], and then developed further in 1995 by Molloy and Reed [16]. In much of this previous work, special emphasis was placed on regular graphs (rather the opposite extreme from what we want here). In 2001 Aiello, Chung, and Lu [1] used Molloy and Reed's model with a power law for the degree sequence. Precisely, they assumed that the number of vertices of degree x is $e^{\alpha}x^{-\beta}$ for some positive constants α and β. They proved results analogous to those in the traditional model. For example, if β is between 2 and 3.47, there is almost surely a unique giant component with $\Theta(n)$ vertices, the second largest component has only $\Theta(\log n)$ vertices, and there are components of essentially every size

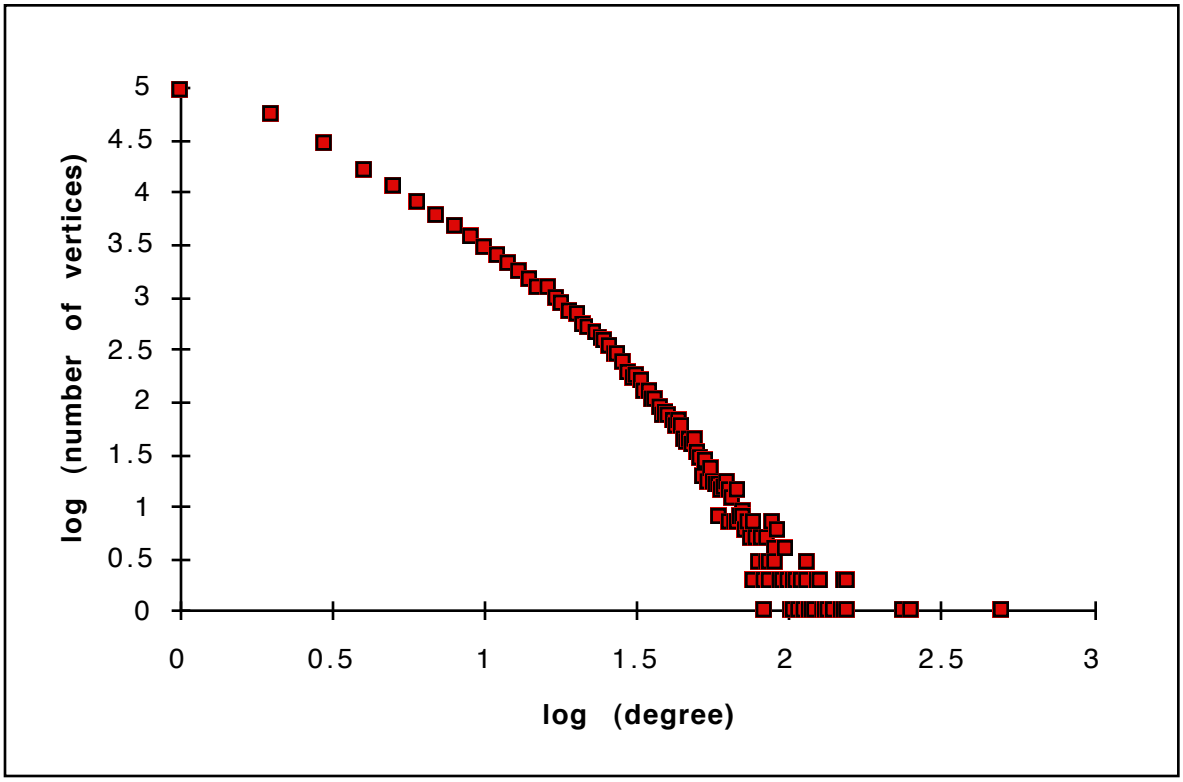

Figure 1. Distribution of (nonzero) vertex degrees in C_{math}

from 2 to $\Theta(\log n)$.

Rather than specify the required degree distribution ahead of time, Barabási and Albert in 1999 [3] grew their graphs with the following algorithm, which incorporates the kinds of assumptions one might want to make about collaboration graphs. Initially, the graph has one vertex (or a small number of vertices). At each time step, a new vertex is added and attached to old ones preferentially, with vertices of higher degree more likely to become neighbors of the new vertices; and, in some variants [5], edges are also added among vertices already present, again with preferential attachment. In simulations, they found that the degrees of the vertices in the resulting graphs did follow a power law, with $\beta \approx 3$. Bollobás et. al [7] confirmed theoretically that $\beta = 3$ in this model.

These are not the only approaches in the literature. Several other models and analyses have been proposed (e.g., a random model without preferential attachment [8]); see [4] for a recent survey, which also explains the connection to percolation theory in statistical mechanics. There is widespread interest in this subject, perhaps partly because of its relevance to analyzing the World Wide Web [14]. Indeed, papers have appeared in such main-stream journals as *The Proceedings of the National Academy of Sciences* [19], *Nature* [24], and *Science* [3], and there are at least two recent books on these topics pitched to a general audience [23], [2].

§4. The Properties of C_{math}

The mathematics research collaboration graph C_{math} has 337,454 vertices and 496,489 edges, so the average degree (number of coauthors for a mathematician) is about 2.94. There are 84,115 isolated vertices in C_{math} (25%), which we should probably 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 3.92.

Let us first look at the degrees of the vertices, and see whether the power law predicted in §3 holds. We want to fit the equation $y = e^{\alpha}x^{-\beta}$ to our data, where x ranges over the degrees of the vertices and y is the number of vertices of degree x. The scatterplot of $\log(y)$ versus $\log(x)$ is shown in Figure 1. (We omit the degrees that do not occur.) A regression gives $\beta = 2.81$, with $R^2 = 93.9\%$. This value of β is consistent with other examples of massive graphs studied in the literature cited in §3, such as the World Wide Web and telephone call graphs. (If we omit Paul Erdős, whose degree is 502, then we have $\beta = 2.87$ and $R^2 = 94.9\%$.)

Next we look at the sizes of the components of C_{math}. One giant component has 208,200 vertices and 461,643 edges; and the remaining 45,139 vertices and 34,846 edges split into 16,883 components, each having from 2 to 39 vertices. The sizes of the nongiant components are shown in Figure 2, whose axes are again on a log-log scale.

According to the models in §3, the component sizes, as well as the vertex degrees, should follow a power law; that is, the number of nongiant components with x vertices should be proportional to x^{δ} for some $\delta < 0$. A linear regression gives us an exponent of $\delta = -3.72$, with $R^2 = 97.3\%$. If we omit the outlier (an accident that the largest nongiant component happens to have 39 vertices as opposed to, say, 25), then we get an even better fit: exponent $\delta = -3.96$ and $R^2 = 99.2\%$. The model in [1] predicts that since β is between 2 and 3.47, we should expect a largest nongiant component to have on the order of $\log(n)$ vertices, and there should be a component of essentially every size up to this value. In fact, C_{math} has components of every size up to 21, and the only two larger nongiant components have 23 and 39 vertices.

Next, we concentrate just on the giant component of C_{math} and consider the distribution of distances between vertices. Based on a random sample of 66 pairs of vertices in this component, the average distance between two vertices is around 7.73 (between 7.37 and 8.08 with 95% confidence), with a standard deviation of about 1.45. 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 buzz phrase for C_{math} may be "nine degrees of separation" [13], if we wish to account for three quarters of all pairs of mathematicians.

We find that the diameter of the giant component (maximum distance between two vertices) is 27, and the radius (minimum eccentricity of a vertex, where the eccentricity is the maximum distance from that vertex to any other) is 14. (Surprisingly, the eccentricity of Paul Erdős is 15, not 14; Noga Alon and at least two other vertices have eccentricity

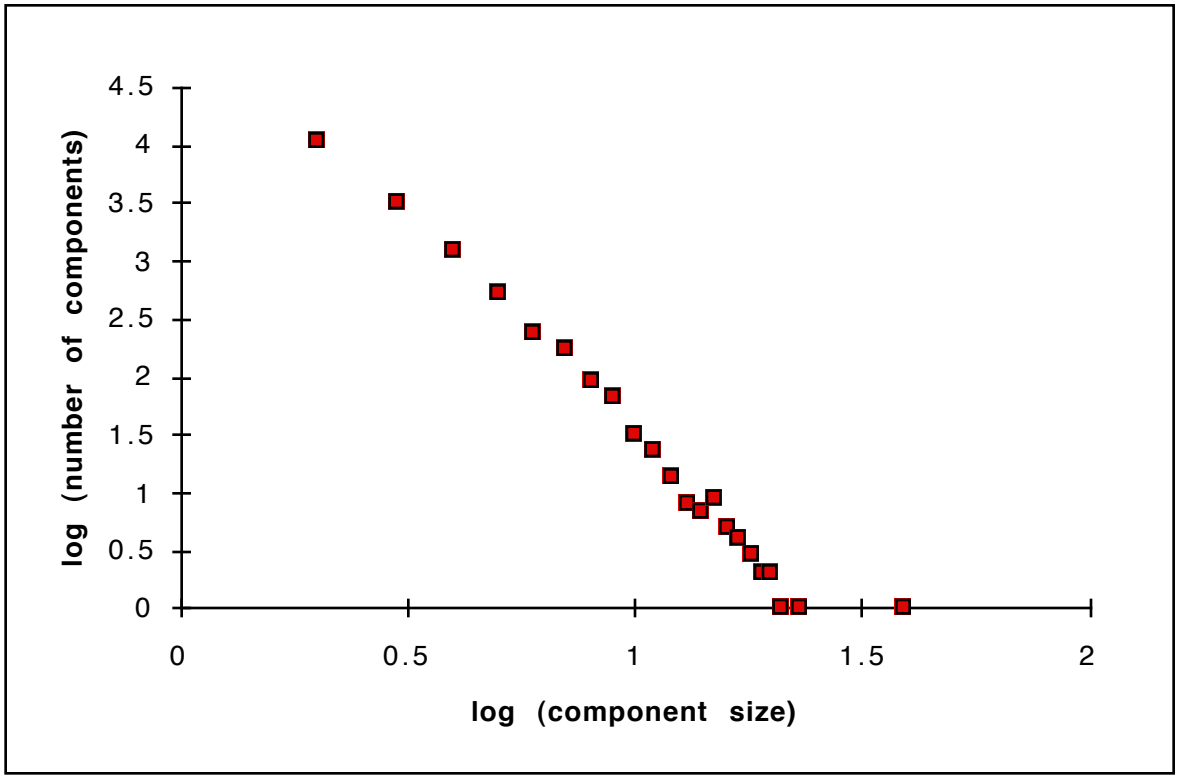

Figure 2. Distribution of (nongiant) component sizes in C_{math}

14.) For an Erdős-Rényi random graph with the edge density of C_{math}, one expects small diameters in the giant component [9]; in our case we should expect something close to $\log(337454)/\log(2.94) \approx 11.8$ (or $\log(253339)/\log(3.92) \approx 9.1$ if we ignore the isolates, or $\log(208200)/\log(4.43) \approx 8.2$ if we stick to the giant component). Theoretical estimates of the radius, diameter, or average path length for other random graph models, however, have not been computed.

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 [12].

To analyze this aspect of C_{math}, we took a random sample of 100 vertices in the large component and computed for each of them the mean and standard deviation of the distances to all the other vertices. The means varies from 5.65 to 11.61, with an average of 7.52 and a standard deviation of 1.00. The standard deviations are remarkably constant, with the numbers varying only between 1.19 and 1.35 (the interquartile range was from 1.23 to 1.27, mean 1.25, standard deviation 0.03). 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 3 shows the distribution of Erdős numbers and the distribution of "Jane Doe" numbers for a person chosen at

random. It seems as if those people further away from the heart of the graph may take longer to get to the heart, but once there, the fan-out pattern is the same.

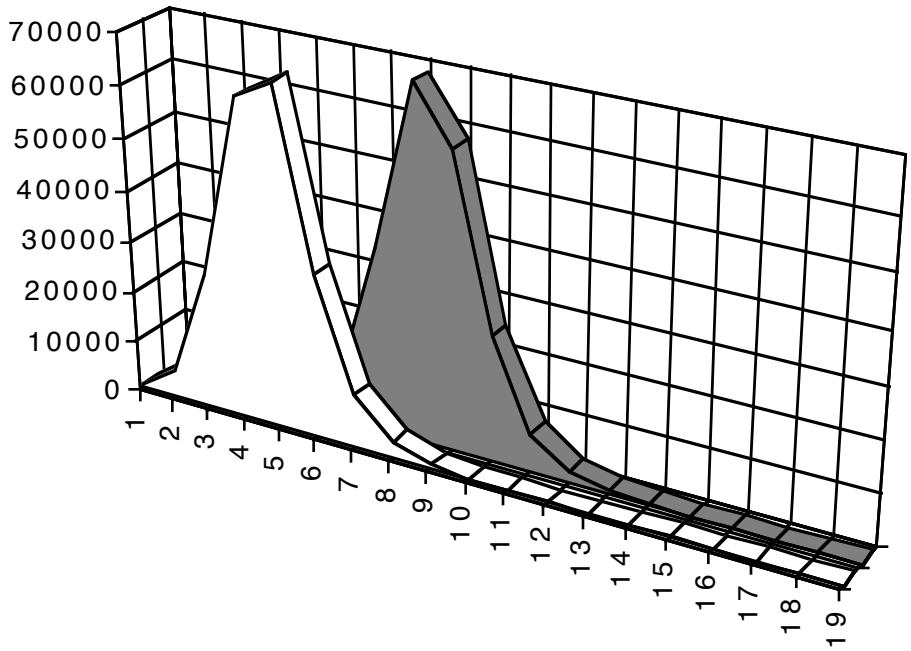

Figure 3. Distribution of Erdős numbers (front) and Doe numbers

Finally, we compute the clustering coefficient of C_{math} to be 0.15 (this calculation is confirmed in [17]). That is 10,000 times higher than one would expect for a traditional random graph with 253,000 vertices and 496,000 edges, another indication of the need for better models [18]. Theoretical estimates of the clustering coefficient for other random graph models, however, have not been computed.

§5. The Evolution of C_{math} over Time

Tables 1 and 2 gives various statistics on the publication habits of mathematicians over time, organized roughly into decades (the 90s end with papers published around 2000). These summaries were determined from B_{math}. All integer figures are rounded to the nearest thousand. Data are given both for all authors, and for authors who have collaborated. A "collaboration" is recorded for each instance of a pair of authors sharing a publication.

These tables reveal a number of trends. The total number of papers per year seems to be rising about 1000–2000 papers per year per year, but the actual increase in publications may be more, to the extent that MR tries to keep its cost in line by becoming more selective in what it catalogs. That the relative rate of increase in the number of authors is slightly less can be seen from the fact that the average number of papers per author per decade has risen slowly from about 4 to about 5 over the past half century.

	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
ave. 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	0.36
1-auth papers	66%	73%	78%	84%	88%	91%
2-auth papers	26%	22%	18%	13%	11%	8%
3-auth papers	7%	4%	3%	2%	1%	1%
$>$ 3-auth papers	1%	1%	1%	1%	0%	0%
ave. 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
collaborating authors	253	153	82	34	11	3
fraction of all authors	75%	68%	60%	49%	39%	28%
ave. collaborators/auth	2.94	2.26	1.67	1.20	0.83	0.49
ave. collaborations/auth	5.65	4.08	2.87	2.01	1.36	0.75
ave. coll'tors/col. auth	3.92	3.33	2.79	2.42	2.14	1.74
ave. coll'tions/col. auth	7.52	6.00	4.77	4.07	3.50	2.65

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

Perhaps most striking is the increasing tendency to collaborate. Whereas in the 1940s and 1950s, only one paper in nine was joint work, and papers with more than two authors were vitually unheard of, by the 1990s only half the papers being published were solo ventures, and one in eight had more than two authors. In those early years, less than half of all mathematicians had ever written a joint paper, whereas four fifths of mathematicians who published in the 1990s had published at least one joint work in that decade. The average number of collaborators for an author has also grown dramatically, from less than one if we close the books in 1959, to nearly three if we look at the entire database.

If we look just at the papers in the late 1990s (the last 95,000 papers in our database, or about two years' worth), then we find that the mean number of authors per paper has increased to 1.71 (standard deviation 1.90), with 49% of the papers by a single author, 35% by two authors, 13% by three authors, and 3% by four or more authors, an increase in collaboration even over the 1990s as a whole.

§6. Open Questions and Directions for Future Work

Using the data in MR, one can look at other questions as well. For example, it would be very interesting to look at the bipartite graph B_{math} itself and study such things as the numbers of papers mathematicians write, and when in their careers they write them; or turn the tables and look at the "collaboration graph" of papers, rather than authors. We can also analyze the subgraphs of C_{math} restricted to various branches of mathematics or specific subjects. In what qualitative and quantitative ways, for example, does

	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
ave. 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-auth papers	54%	66%	73%	81%	87%	91%
2-auth papers	33%	27%	22%	16%	11%	8%
3-auth papers	10%	6%	4%	2%	2%	1%
$>$ 3-auth papers	3%	1%	1%	1%	0%	0%
ave. 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
collaborating authors	155	104	62	27	9	3
fraction of all authors	81%	72%	64%	52%	41%	28%
ave. collaborators/auth	2.84	2.16	1.62	1.18	0.84	0.49
ave. collaborations/auth	5.14	3.66	2.62	1.90	1.34	0.75
ave. coll'tors/col. auth	3.51	2.99	2.55	2.25	2.08	1.74
ave. coll'tions/col. auth	6.35	5.06	4.11	3.64	3.31	2.65

Table 2. Data for each decade; integer figures in thousands

the subgraph C_{05} of C_{math}, in which only those papers are considered whose primary mathematics subject classification [22] is "05 Combinatorics" (which includes enumerative combinatorics, designs and configurations, graph theory, extremal combinatorics, and algebraic combinatorics), differ from all of C_{math}? We conjecture that the publishing and collaboration habits strongly depend on subfield.

Acknowledgment

The author thanks the American Mathematical Society for providing access to the data needed for the analyses presented in this paper, as well as Patrick Ion of *Mathematical Reviews* for helpful conversations.

References

1. William Aiello, Fan Chung, and Linyuan Lu, A random graph model for power law graphs, *Experimental Mathematics* **10** (2001) 53–66; **MR** 2001m:05233.

2. Albert-László Barabási, *Linked: The New Science of Networks*, Perseus, 2002.

3. Albert-László Barabási and Réka Albert, Emergence of scaling in random networks, *Science* **286** (1999) 509–512.

4. Albert-László Barabási and Réka Albert, Statistical mechanics of complex networks, *Reviews of Modern Physics* **74** (2002) 47–97.

5. Albert-László Barabási, Réka Albert, and Hawoong Jeong, Scale-free characteristics of random networks: the topology of the world-wide web, *Physica A* **281** (2000) 69–77.

6. Edward A. Bender and Rodney E. Canfield, The asymptotic number of labeled graphs with given degree sequences, *J. Combinat. Theory Ser. A* **24** (1978) 296–307; **MR** 58#21793.

7. Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády, The degree sequence of a scale-free random graph process, *Random Structures & Algorithms* **18** (2001) 279–290; **MR** 2002b:05121.

8. Duncan S. Callaway, John E. Hopcroft, Jon M. Kleinberg, M. E. J. Newman, and Steven H. Strogatz, Are randomly grown graphs really random?, *Physical Review E* **64** (2001) 041902.

9. Fan Chung and Linyuan Lu, The diameter of sparse random graphs, *Advances in Applied Mathematics* **26** (2001) 257–279; **MR** 2002c:05138.

10. 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.

11. Jerrold W. Grossman, The Erdős Number Project, `http://www.oakland.edu/~grossman/erdoshp.html`.

12. Jerrold W. Grossman and Patrick 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, FL, 1995), *Congressus Numerantium* **108** (1995) 129–131; **CMP** 1 369 281.

13. John Guare, *Six Degrees of Separation*, Random House, 1990.

14. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal, Stochastic models for the Web graph, *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science* (FOCS 2000) 57–65.

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

16. Michael Molloy and Bruce Reed, A critical point for random graphs with a given degree sequence, *Random Structures & Algorithms* **6** (1995) 161–179; **MR** 97a:05191.

17. M. E. J. Newman, Ego-centered networks and the ripple effect, or Why all your friends are weird, preprint, `http://arXiv.org/pdf/cond-mat/0111070`.

18. M. E. J. Newman, Random graphs as models of networks, preprint, `http://arXiv.org/pdf/cond-mat/0202208`.

19. 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.

20. M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications, *Physical Review E* **64** (2001) 026118.

21. 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`.

22. 2000 Mathematics Subject Classification, American Mathematical Society, `http://www.ams.org/msc`.

23. Duncan J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, 1999; **MR** 2001a:91064.

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