On a Portion of the Well-Known Collaboration Graph

Jerrold W. Grossman
Oakland University
Rochester, MI 48309-4401
e-mail: grossman@oakland.edu

Patrick D. F. Ion
Mathematical Reviews
Ann Arbor, MI 48107-8604
e-mail: ion@math.ams.org

The following definitions are probably well-known among many research mathematicians. The *collaboration graph* C has as vertices all researchers (mathematicians, in particular), with an edge joining u and v if u and v have published a joint research paper. Although there are different possibilities for handling papers with more than two authors (the best approach probably being to use hypergraphs), we will simply put a single edge between every pair of authors of a multi-author paper. (A more detailed treatment would also use parallel edges to represent repeated collaboration.) A distinguished vertex p of C is the outstanding, prolific, and venerable Paul Erdős. The distance from a vertex u to p is known as u's *Erdős number*. Thus, for example, Paul Erdős's co-authors have Erdős number 1. Those people with finite Erdős number constitute the *Erdős component* of C. This note reports on an on-going project in which a portion of this graph—in particular, a list of all people with small Erdős numbers—is made available in electronic form. We discuss the difficulties of obtaining sound data, state a few interesting properties of this portion of the graph, issue a plea to authors, and list some open questions.

The most recent and extensive study of C to date was done by Ron Graham [3], but it does not aim to be complete. We should also point out that portions of our data can provide fairly large, interesting "real-life" graphs on which to test graph algorithms, in the spirit of [1]. Since the data change over time, both because of additional collaborations and because of corrections to inaccurate information, we intend to update the lists on a regular basis and label each version. It is intended that these lists will be available electronically via anonymous FTP and possibly on the World Wide Web; interested readers should contact the first-named author for details.

The primary source of information for this project is various data bases of publications in the mathematical sciences, notably those maintained by the American Mathematical Society's *Mathematical Reviews* (*MR*). *MR* has published short reviews of essentially all mathematical research books and articles since 1940 (including many works in statistics and computer science). Only a few thousand items were reviewed per year fifty years ago, whereas now the rate is around 50,000. More interesting than sheer volume is the proportion of items that are the result of collaboration. Figure 1 shows the fraction of all authored items in *MR* with one, two, or more than two authors, as a function of time. Notice that while over 90% of all papers fifty years ago were the work of just one mathematician, today scarcely more than half of them are solo works. In the same period, the fraction of two-author papers has risen from under 10% to about one third. Also, in 1940 there were virtually no papers with three authors, let alone four or more; now about 10% of all papers in the mathematical sciences have three or more authors, including about 2% with four or more.

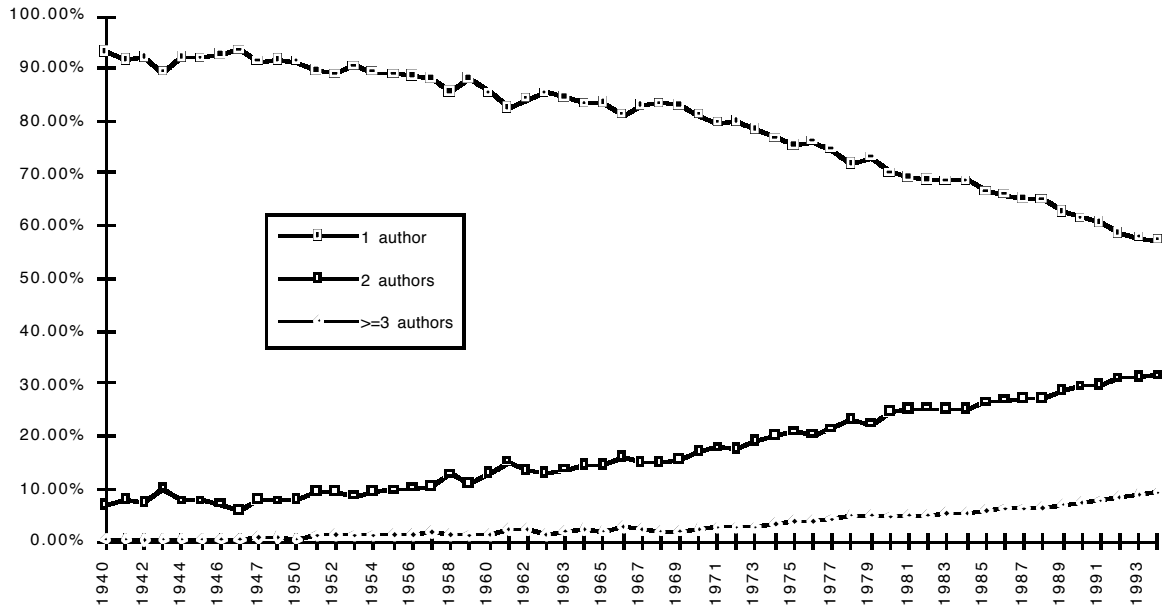

Figure 1. The percentages of papers in *MR* with one, two, or more authors.

One can speculate on the reasons for this trend, such as increases in mathematics department size, the growth and specialization of the field, proliferation of scholarly meetings, electronic communication, and of course the "publish or perish" pressure on faculty. We suspect that the distribution of number of authors is also not constant within the mathematical science, from subfield to subfield. For example, the papers in this conference proceedings (dealing with discrete mathematics) will likely show a larger number of authors than the current averages across all of mathematics. (Of course, broad collaboration is much more common in the laboratory sciences, with an average of about 3.5 authors per paper [2]. The Institute for Scientific Information found 407 papers published in 1994 with more than 50 authors, including 18 with more than 500; the record seems to be 972.)

In compiling data such as ours, one must face the serious problem of identifying authors on the basis of name strings appearing on papers. For example, *MR* knows of two persons publishing under the name "Paul Erdös"—the vertex p mentioned above (who uses a wide variety of name strings, including "P. Erdős") and a Hungarian physicist. Fortunately, Peter L. Erdös is usually not confused with p because he uses his full first name and/or his middle initial. As another example, there are two distinct persons named Norman Lloyd Johnson, each publishing under the same four name strings; one is a geometer from Iowa with over 90 articles in the past dozen years, the other a statistician from North Carolina with over 30 articles in the same time period. It is inevitable that the lists we have compiled contain mistakes of identity, as well as omissions and other errors, and we would be most happy to have all of them brought to our attention. We issue a much more substantial plea, however: *every author should pick as complete an author name as possible for his or her publications and use it on every item.* As a corollary, references to other people and their work should use this preferred and complete name string.

The data that we have compiled as of this conference are as follows. First, we

have a list of the approximately 450 co-authors of Paul Erdős, including the date of their first joint work (the first being a 1934 paper by Erdős and George Szekeres, and the most recent published in 1994). Second, for each of these people with Erdős number 1, we have a list of their co-authors. This list includes other people with Erdős number 1 as well as the approximately 4300 people with Erdős number 2.

Let E_1 be the subgraph of the collaboration graph (as of January 1995) induced by people with Erdős number 1. We found that E_1 has 451 vertices and 1145 edges. Furthermore, these collaborators tended to collaborate a lot, especially among themselves. They have an average of 19 other collaborators (standard deviation 21), and only seven of them collaborated with no one except Erdős. Four of them have over 100 co-authors. If we restrict our attention just to E_1, we still find a lot of joint work. Only 41 of these 451 people have collaborated with no persons having Erdős number 1 (i.e., there are 41 isolated vertices in E_1), and E_1 has four components with two vertices each. The remaining 402 vertices in E_1 induce a connected subgraph. The average vertex degree in E_1 is 5 (standard deviation 6); and there are four vertices with degrees of 30 or higher. The largest clique in E_1 has seven vertices, but it should be noted that six of these people and Erdős have a joint seven-author paper. In addition, there are seven maximal 6-cliques and 61 maximal 5-cliques. In all, 29 vertices in E_1 are involved in cliques of order 5 or larger. Finally, we found (by computer) that the diameter of E_1 is 11 and its radius is 6.

Three quarters of the people with Erdős number 2 have only one co-author with Erdős number 1 (i.e., each such person has a unique path to p of length 2). However, their mean number of Erdős number 1 co-authors is 1.5 (standard deviation 1.1), and the count ranges as high as 13.

Folklore has it that most active researchers have a finite, and fairly small, Erdős number. For supporting evidence, we verified that all the Fields and Nevanlinna prize winners during the past three cycles (1986–1994) are indeed in the Erdős component, with Erdős number at most 9. Since this group includes people working in theoretical physics, one can conjecture that most physicists are also in the Erdős component, as are, therefore, most scientists in general. The large number of applications of graph theory and statistics to the social sciences might also lead one to suspect that many researchers in other academic areas are included as well. We close with two open questions about C, restricted to mathematicians, that such musings suggest, with no hope that either will ever be answered satisfactorily: What is the diameter of the Erdős component, and what is the order of the second largest component?

References

[1] Donald Knuth, *The Stanford GraphBase* (Addison-Wesley, 1993)

[2] Kim A. McDonald, Too many co-authors?, *Chronicle of Higher Education* (April 28, 1995) A35–A36

[3] Tom Odda, On properties of a well-known graph or what is your Ramsey number?, *Topics in Graph Theory* (New York, 1977) 166–172