If People Were Papers and Papers Were People

Jerrold W. Grossman
Oakland University
Rochester, Michigan

AMS Annual Meeting
January 10, 2004
Phoenix, Arizona

Let $\boldsymbol{G}$ be a bipartite graph, with vertex sets $\boldsymbol{A}$ (authors) and $\boldsymbol{P}$ (papers) — e.g., the "authorship graph" for mathematics papers.

Form the author collaboration graph C_A on vertex set A, in which two vertices are adjacent if they have a common neighbor in P.

Similarly, form C_P, the "paper collaboration graph" on P.

People who have studied actual social networks such as C_A (Albert-László Barabási, Fan Chung, J. G., Mark Newman, …), as well as similar large graphs such as power grids, the Internet, and neural networks, have found that in such graphs:

There is a giant component containing most of the vertices.

The diameter of the giant component is fairly small.

The clustering coefficient is relatively high.

The vertex degrees follow a power law distribution.

The Data

All authored items from
Mathematical Reviews
(MathSciNet)

1940–1999

MR tries hard to identify authors as people,
not name strings.

340,000 authors (***A***)

1,600,000 papers (***P***)

2,300,000 edges

Facts About the Underlying Bipartite Graph

(the Authorship Graph)

	authors	papers
mean degree	6.87	1.45
median degree	2	1
max degree	1500	15?

Facts About the Author and Paper Collaboration Graphs, C_A and C_P

	authors	**papers**
# vertices	337,000	1,598,000
# edges	496,000	44,685,000
mean degree	2.94	56
max degree	502	2090

Ignoring the isolated vertices:

# vertices	253,000	1,541,000
mean degree	3.92	58
median degree	2	33

giant component	62%	82%
isolated vertices	25%	4%
other vertices	13%	14%

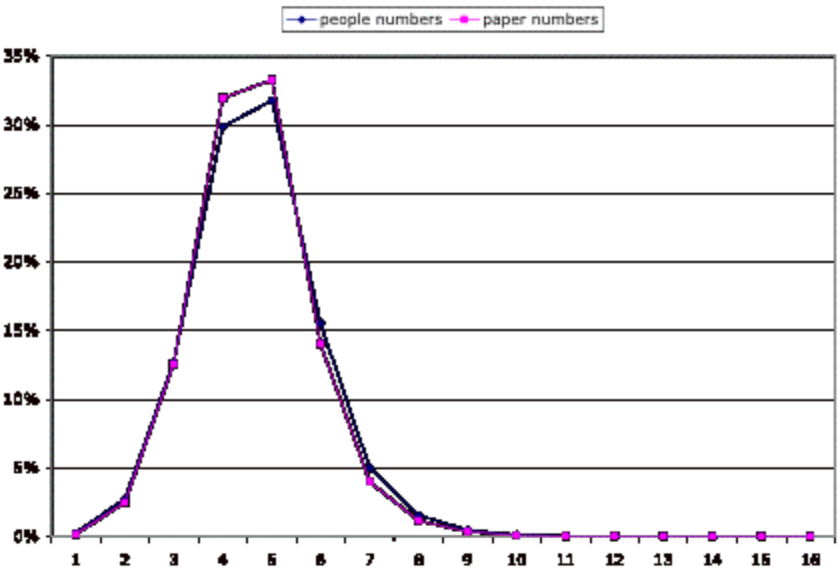
Degree distribution for non-isolated vertices:

	authors	papers
degree 1	37%	2.9%
degree 2	22%	2.6%
degree 3	12%	2.4%
degree 4	7%	2.2%
degrees 5–10	14%	12%
degrees > 10	8%	78%

Distribution of Distances in (the Giant Components of) C_A and C_P

Diameters are essentially the same, because the distance between two papers in C_P is within 1 of the distance between their authors in C_A. Diameter is around 28.

Pick a "center" vertex and call it e (for Paul Erdös). The Erdös number of vertex v (with respect to vertex e) is the distance from v to e. What is the distribution of Erdös numbers in C_A and C_P?


people numbers
paper numbers
35%
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Some Graph Theory Questions

1. Given a desired C_A, what is the minimum number of papers needed in G (elements of P) that will produce it?

The edge clique cover number of a graph H is the smallest number of complete subgraphs of H whose union includes all the edges of H. The intersection number of H is the minimum cardinality of a set S such that H is the intersection graph of a collection of subsets of S. These two values are the same and give the answer to this question.

(But it's NP-hard to compute.)

2. Given an authorship graph G, what is the minimum number of papers of G that will produce the same author collaboration graph C_A? In other words, what fraction of the papers can be thrown away without changing the author collaboration graph?

For more information, visit the Erdös Number Project website:

www.oakland.edu/~grossman/erdoshp.html

www.oakland.edu /~grossman/erdoshp.html