Network Modeling for Citation Analysis

BINF739 SPRING2007

Jeff Solka and Jennifer Weller

References

- M. E. J. Newman, "Who is the best connected scientist? A study of scientific coauthorship networks," Phys. Rev. E64 (2001).

Affiliation Networks

- An affiliation network is a network of actors connected by common membership in groups of some sort, such as clubs, teams, or organizations.
  - Women and the social
  - Company CEOs and the clubs they frequent [25]
  - Company directors and the boards of directors on which they sit
  - Movie actors the movies in which they appear
- Networks tend to be more reliable than those on other social networks, since membership of a group can often be determined with a precision not available when considering friendship or other types of acquaintance.

Coauthorship Affiliation Networks

- An affiliation networks of scientists in which a link between two scientists is established by their co-authorship of one or more scientific papers.
- Thus the groups to which scientists belong in this network the groups of coauthors of a single paper.
- This network is in some ways more truly a social network than many affiliation networks; it is probably fair to say that most pairs of people who have written a paper together are genuinely acquainted with one another, in a way that movie actors who appeared together in a movie may not be.
- There are exceptions—some very large collaborations, for example in high-energy physics, will contain coauthors who have never even met—and we will discuss these at the appropriate point.
- Similar concerns arise in medical or bioinformatics type papers.

Erdos Number

- Based on your "degree of separation" from Paul Erdos.
- If you had published a paper with Paul Erdos you would have an Erdos number of 1, if you had published a paper with someone who had published a paper with him you would have an Edrod number of 2, etc.
- Your Erdos number is the geodesic distance between you and Erdos in the coauthorship network.
- The average Erdos number is about 4.7, and the maximum known finite Erdos number (within mathematics) is 15.

Data Sources

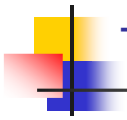
- Los Alamos e-Print Archive
  - A database of unrefereed preprints in physics, self-submitted by their authors, running from 1992 to the present.
- MEDLINE
- SPIRES
  - A database of preprints and published papers in high-energy physics
- NCSTRL
  - A database of preprints in computer science

Focus of Study

- Five year period from 1995 to 1999
  - First, older data are less complete than newer for all databases.
  - Second, we wanted to study the same time period for all databases, so as to be able to make valid comparisons between collaboration patterns in different fields.

Building the Network

- Read through paper list keep track of existing nodes and add in edges as needed.
  - O(pn)
- Binary tree implementation
  - O(p * log(n))



The Alias Problem - I

- The size of the databases varies considerably from about a million authors for MEDLINE to about ten thousand for NCSTRL.
- In fact, it is difficult to say with precision how many authors there are. One can say how many distinct names appear in a database, but the number of names is not the same as the number of authors.
- A single author may report their name differently on different papers. For example, F. L. Wright, Francis Wright, and Frank Lloyd Wright could all be the same person.
- Also two authors may have the same name. Grossman and Ion [39] point out that there are two American mathematicians named Norman Lloyd Johnson, who are known to be distinct people and who work in different fields, but between whom computer programs such as ours cannot hope to distinguish.
- Even additional clues such as home institution or field of specialization cannot be used to distinguish such people, since many scientists have more than one institution or publish in more than one field.

The Alias Problem - II

- In the first, we identify each author by his or her surname and first initial only.
  - This method is clearly prone to confusing two people for one, but will rarely fail to identify two names which genuinely refer to the same person.
- In the second version of each network, we identify authors by surname and all initials.
  - This method can much more reliably distinguish authors from one another, but will also identify one person as two if they give their initials differently on different papers.
  - Indeed this second measure appears to overestimate the number of authors in a database substantially.

Summary Results

	MEDLINE	Los Alamos e-Print Archive				SPIRES	NCSTRL
		complete	astro-ph	cond-mat	hep-th		
total papers	2163923	98502	22029	22016	19085	66652	13169
total authors	1520251	52909	16706	16726	8361	56627	11994
first initial only	1090584	45685	14303	15451	7676	47445	10998
mean papers per author	6.4(6)	5.1(2)	4.8(2)	3.65(7)	4.8(1)	11.6(5)	2.55(5)
mean authors per paper	3.754(2)	2.530(7)	3.35(2)	2.66(1)	1.99(1)	8.96(18)	2.22(1)
collaborators per author	18.1(1.3)	9.7(2)	15.1(3)	5.86(9)	3.87(5)	173(6)	3.59(5)
size of giant component	1395693	44337	14845	13861	5835	49002	6396
first initial only	1019418	39709	12874	13324	5593	43089	6706
as a percentage	92.6(4)%	85.4(8)%	89.4(3)	84.6(8)%	71.4(8)%	88.7(1.1)%	57.2(1.9)%
2nd largest component	49	18	19	16	24	69	42
clustering coefficient	0.066(7)	0.43(1)	0.414(6)	0.348(6)	0.327(2)	0.726(8)	0.496(6)
mean distance	4.6(2)	5.9(2)	4.66(7)	6.4(1)	6.91(6)	4.0(1)	9.7(4)
maximum distance	24	20	14	18	19	19	31

TABLE I. Summary of results of the analysis of seven scientific collaboration networks. Numbers in parentheses are standard errors on the least significant figures.

Number of Papers Per Author - I

- The average number of papers per author in the various subject areas is in the range of around three to six over the five year period.
- The only exception is the SPIRES database, covering high-energy physics, in which the figure is significantly higher at 11.6.
  - One possible explanation for this is that SPIRES is the only database which contains both preprints and published papers. It is possible that the high figure for papers per author reflects duplication of papers in both preprint and published form.

Number of Papers Per Author - II

- In addition to the average numbers of papers per author in each database, it is interesting to look at the distribution pk of numbers k of papers per author.
- In 1926, Alfred Lotka [50] showed, using a dataset compiled by hand, that this distribution followed a power law, with exponent approximately −2, a result which is now referred to as Lotka's Law of Scientific Productivity.
- In other words, in addition to the many authors who publish only a small number of papers, one expects to see a "fat tail" consisting of a small number of authors who publish a very large number of papers.

Log Log Plot of the Number of Paper Histograms

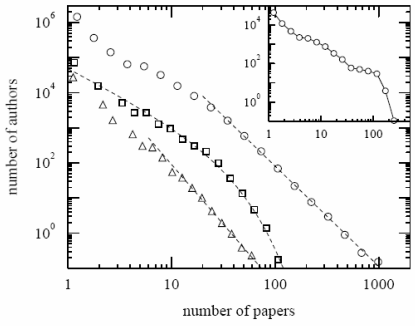


FIG. 1. Histograms of the number of papers written by authors in MEDLINE (circles), the Los Alamos Archive (squares), and NCSTRL (triangles). The dotted lines are fits to the data as described in the text. Inset: the equivalent histogram for the SPIRES database.

(These histograms and all the others shown here were created using the "all initials" versions of the collaboration networks.)

Power Law Fits to the Data - I

- For the MEDLINE and NCSTRL databases these histograms follow a power law quite closely, at least in their tails, with exponents of −2.86(3) and −3.41(7) respectively—somewhat steeper than those found by Lotka, but in reasonable agreement with other more recent studies.

Power Law Fits to the Data - II

- For the Los Alamos Archive the pure power law is a poor fit.
- An exponentially truncated power law does much better:

$$p_k = Ck^{-\tau} e^{-k/\kappa}, \qquad (1)$$

where τ and κ are constants and C is fixed by the requirement of normalization—see Fig. 1. (The probability p_0 of having zero papers is taken to be zero, since the names of scientists who have not written any papers do not appear in the database.) The exponential cutoff we attribute to the finite time window of five years used in this study which prevents any one author from publishing a very large number of papers. Lotka and subsequent authors who have confirmed his law have not usually used such a window.

The Exponential Cutoff or Lack Thereof

- Why does the cutoff appear in physics and not computer science or biomedicine?
- Thus it is possible that there is not after all any fat tail in the distribution for the MEDLINE database, only the illusion of one produced by the large number of scientists with commonly occurring names.
- For the SPIRES database, which is shown separately in the inset of the figure, neither pure nor truncated power law fits the data well, the histogram displaying a significant bump around the 100-paper mark.
- A possible explanation for this is that a small number of large collaborations published around this number of papers during the time-period studied.
- Since each author in such a collaboration is then credited with publishing a hundred papers, the statistics in the tail of the distribution

Authors With the Highest Numbers of Papers

	number of papers		number of co-workers		betweenness ($\times 10^6$)		collaboration weight	
astro-ph	112	Fabian, A.C.	360	Frontera, F.	2.33	Kouveliotou, C.	16.5	Moskalenko, I.V./Strong, A.W.
	101	van Paradijs, J.	353	Kouveliotou, C.	2.15	van Paradijs, J.	15.0	Hernquist, L./Heyl, J.S.
	81	Frontera, F.	329	van Paradijs, J.	1.80	Filippenko, A.V.	14.0	Mathews, W.G./Brighenti, F.
	80	Hernquist, L.	299	Piro, L.	1.57	Beaulieu, J.P.	13.4	Labini, F.S./Pietronero, L.
	79	Gould, A.	296	Costa, E.	1.52	Nomoto, K.	12.2	Piran, T./Sari, R.
	78	Silk, J.	291	Feroci, M.	1.52	Pian, E.	11.8	Zaldarriaga, M./Seljak, U.
	78	Klis, M.V.D.	284	Pian, E.	1.49	Frontera, F.	11.4	Hernquist, L./Katz, N.
	73	Kouveliotou, C.	284	Hurley, K.	1.35	Silk, J.	11.1	Avila-Reese, V./Firmani, C.
	70	Ghisellini, G.	244	Palazzi, E.	1.33	Kamionkowski, M.	10.9	Dai, Z.G./Lu, T.
	66	Piro, L.	244	Heise, J.	1.28	McMahon, R.G.	10.8	Ostriker, J.P./Cen, R.
cond-mat	116	Parisi, G.	107	Uchida, S.	4.11	MacDonald, A.H.	22.3	Belitz, D./Kirkpatrick, T.R.
	79	Scheffler, M.	103	Ueda, Y.	3.96	Bishop, A.R.	17.0	Shrock, R./Tsai, S.
	75	Das Sarma, S.	96	Revcolevschi, A.	3.36	Das Sarma, S.	15.0	Yukalov, V.I./Yukalova, E.P.
	74	Stanley, H.E.	94	Eisaki, H.	2.96	Tosatti, E.	14.7	Martín-Delgado, M.A./Sierra, G.
	70	MacDonald, A.H.	84	Cheong, S.	2.52	Wang, X.	14.3	Krapivsky, P.L./Ben-Naim, E.
	68	Sornette, D.	83	Isobe, M.	2.38	Revcolevschi, A.	14.1	Beenakker, C.W.J./Brouwer, P.W.
	60	Volovik, G.E.	78	Stanley, H.E.	2.30	Uchida, S.	13.8	Weng, Z.Y./Sheng, D.N.
	56	Beenakker, C.W.J.	76	Shirane, G.	2.21	Sigrist, M.	13.7	Sornette, D./Johansen, A.
	53	Dagotto, E.	76	Scheffler, M.	2.19	Cheong, S.	13.6	Rikvold, P.A./Novotny, M.A.
	50	Helbing, D.	76	Menovsky, A.A.	2.18	Stanley, H.E.	13.0	Scalapino, D.J./White, S.R.
hep-th	78	Odintsov, S.D.	50	Ambjorn, J.	0.98	Odintsov, S.D.	34.0	Lu, H./Pope, C.N.
	73	Lu, H.	44	Ferrara, S.	0.88	Ambjorn, J.	29.0	Odintsov, S.D./Nojiri, S.
	72	Pope, C.N.	43	Vafa, C.	0.88	Kogan, I.I.	18.7	Lee, H.W./Myung, Y.S.
	69	Cvetic, M.	39	Odintsov, S.D.	0.84	Henneaux, M.	18.3	Schweigert, C./Fuchs, J.
	68	Ferrara, S.	39	Kogan, I.I.	0.73	Douglas, M.R.	14.7	Ovrut, B.A./Waldram, D.
	65	Vafa, C.	36	Proeyen, A.V.	0.67	Ferrara, S.	14.7	Kleihaus, B./Kunz, J.
	65	Tseytlin, A.A.	35	Fre, P.	0.63	Vafa, C.	12.9	Mavromatos, N.E./Ellis, J.
	65	Mavromatos, N.E.	35	Ellis, J.	0.60	Khare, A.	12.4	Kachru, S./Silverstein, E.
	63	Witten, E.	35	Douglas, M.R.	0.58	Tseytlin, A.A.	11.7	Kakushadze, Z./Tye, S.H.H.
	54	Townsend, P.K.	34	Lu, H.	0.58	Townsend, P.K.	11.6	Arefeva, I.Y./Volovich, I.V.

TABLE II. The authors with the highest numbers of papers, numbers of coauthors, and betweenness, and strongest collaborations in astrophysics, condensed matter physics, and high-energy theory. The figures for betweenness have been divided by 10^6. Full lists of the rankings of all the authors in these databases can be found on the world-wide web [23].

The Number of Authors Per Paper

- Grossman and Ion report that the average number of authors on papers in mathematics has increased steadily over the last sixty years, from a little over 1 to its current value of about 1.5.
- Higher numbers still seem to apply to current studies in the sciences. Purely theoretical papers appear to be typically the work of two scientists, with high-energy theory and computer science showing averages of 1.99 and 2.22 in our calculations.
- For databases covering experimental or partly experimental subject areas the averages are, not surprisingly, higher:
  - 3.75 for biomedicine
  - 3.35 for astrophysics
  - 2.66 for condensed matter physics
- The SPIRES high-energy physics database however shows the most startling results, with an average of 8.96 authors per paper, obviously a result of the presence of papers in the database written by very large collaborations.
- (Perhaps what is most surprising about this result is actually how small it is. The hundreds strong mega-collaborations of CERN and Fermilab are sufficiently diluted by theoretical and smaller experimental groups, that the number is only 9, and not 100.)

Log Log Histogram of the Number of Authors Per Paper

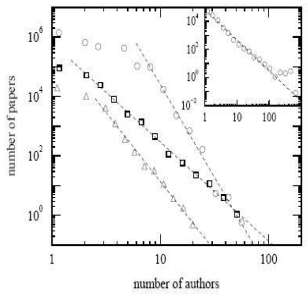


FIG. 2. Histograms of the number of authors on papers in MEDLINE (circles), the Los Alamos Archive (squares), and NCSTRL (triangles). The dotted lines are the best fit power-law forms. Inset: the equivalent histogram for the SPIRES database, showing a clear peak in the 200 to 500 author range.

Distributions of numbers of authors per paper are shown in Fig. 2, and appear to have power-law tails with widely varying exponents of −6.2(3) (MEDLINE), −3.34(5) (Los Alamos Archive), −4.6(1) (NCSTRL), and −2.18(7) (SPIRES). The SPIRES data, which are again shown in a separate inset, also display a pronounced peak in the distribution around 200–500 authors. This peak presumably corresponds to the large experimental collaborations which dominate the upper end of this histogram.

Number of Collaborators Per Author

- The differences between the various disciplines represented in the databases are emphasized still more by the numbers of collaborators that a scientist has, the total number of people with whom a scientist wrote papers during the five-year period.
- The average number of collaborators is markedly lower in the purely theoretical disciplines (3.87 in high-energy theory, 3.59 in computer science) than in the wholly or partly experimental ones (18.1 in biomedicine, 15.1 in astrophysics).
- But the SPIRES high-energy physics database takes the prize once again, with scientists having an impressive 173 collaborators, on average, over a five-year period.
- This clearly begs the question whether the high-energy coauthorship network can be considered an accurate representation of the high energy physics community at all; it seems unlikely that an author could know 173 colleagues well.

Log Log Histogram of the Number of Collaborators

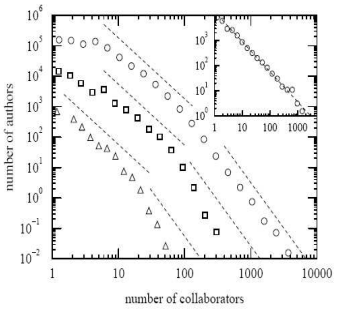
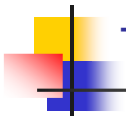


FIG. 3. Histograms of the number of collaborators of authors in MEDLINE (circles), the Los Alamos Archive (squares), and NCSTRL (triangles). The dotted lines show how power-law distributions with exponents −2 and −3 would look on the same axes. Inset: the equivalent histogram for the SPIRES database, which is well fit by a single power law (dotted line).

- The distributions of numbers of collaborators are shown in Fig. 3.
- In all cases they appear to have long tails, but only the SPIRES data (inset) fit a power-law distribution well, with a low measured exponent of −1.20.
- Note also the small peak in the SPIRES data around 700—presumably again a result of the presence of large collaborations.



The Model

- An alternative possibility has been suggested by Barab´asi, based on models of the collaboration process.
- In one such model [48], the distribution of the number of collaborators of an author follows a power law with slope −2 initially, changing to slope −3 in the tail, the position of the crossover depending on the length of time for which the collaboration network has been evolving.
- We show slopes −2 and −3 as dotted lines on the figure, and the agreement with the curvature seen in the data is moderately good, particularly for the MEDLINE data. (For the Los Alamos and NCSTRL databases, the slope in the tail seems to be somewhat steeper than −3.)

Number of Collaborators Per Author

- Column 2 of Table II shows the authors in astro-ph, cond-mat, and hep-th with the largest numbers of collaborators.
- The winners in this race tend to be experimentalists, who conduct research in larger groups, though there are exceptions.
- The high-energy theory database of course contains only theorists, and the smaller numbers of collaborators reflects this.

Size of the Giant Component - I

- In the theory of random graphs it is known that there is a continuous phase transition with increasing density of edges in a graph at which a "giant component" forms, i.e., a connected subset of vertices whose size scales extensively. Well above this transition, in the region where the giant component exists, the giant component usually fills a large portion of the graph, and all other components (i.e., connected subsets of vertices) are small, typically of size O(log n), where n is the total number of vertices.

Size of the Giant Component - II

- We see a situation reminiscent of this in all of the graphs studied here: a single large component of connected vertices which fills the majority of the volume of the graph, and a number of much smaller components filling the rest. In Table I we show the size of the giant component for each of our databases, both as total number of vertices and as a fraction of system size.
- In all cases the giant component fills around 80% or 90% of the total volume, except for high-energy theory and computer science, which give smaller figures.
- A possible explanation of these two anomalies may be that the corresponding databases give poorer coverage of their subjects. The hep-th high-energy database is quite widely used in the field, but overlaps to a large extent with the longer established SPIRES database, and it is possible that some authors neglect it for this reason
- The NCSTRL computer science database differs from the others in this study in that the preprints it contains are submitted by participating institutions, of which there are about 160. Preprints from institutions not participating are mostly left out of the database, and its coverage of the subject area is, as a result, incomplete.

Size of the Giant Component

- The figure of 80–90% for the size of the giant component is a promising one. It indicates that the vast majority of scientists are connected via collaboration, and hence via personal contact, with the rest of their field.
- Despite the prevalence of journal publishing and conferences in the sciences, person-to-person contact is still of paramount importance in the communication of scientific information, and it is reasonable to suppose that the scientific enterprise would be significantly hindered if scientists were not so well connected to one another.

Clustering Coefficients - I

- An interesting idea circulating in social network theory currently is that of "transitivity," which, along with its sibling "structural balance," describes symmetry of interaction amongst trios of actors.
- "Transitivity" has a different meaning in sociology from its meaning in mathematics and physics, although the two are related. It refers to the extent to which the existence of ties between actors A and B and between actors B and C implies a tie between A and C.
- The transitivity, or more precisely the fraction of transitive triples, is that fraction of connected triples of vertices which also form "triangles" of interaction.
- Here a connected triple means an actor who is connected to two others. In the physics literature, this quantity is usually called the clustering coefficient C.

Clustering Coefficients - II

- The factor of three in the numerator compensates for the fact that each complete triangle of three vertices contributes three connected triples, one centered on each of the three vertices, and ensures that C = 1 on a completely connected graph.
- On all unipartite random graphs C = O(n−1) [4,23], where n is the number of vertices, and hence goes to zero in the limit of large graph size.
- In social networks it is believed that the clustering coefficient will take a non-zero value even in very large networks, because there is a finite (and probably quite large) probability that two people will be acquainted if they have another acquaintance in common.
- This is a hypothesis we can test with our collaboration networks. In Table I we show values of the clustering coefficient C, calculated from Eq. (2), for each of the databases studied, and as we see, the values are indeed large, as large as 0.7 in the case of the SPIRES database, and around 0.3 or 0.4 for most of the others.

$$C = \frac{3\times \text{ number of triangles on the graph}}{\text{number of connected triples of vertices}}. \quad (2)$$

Clustering Coefficients - III

- There are a number of possible explanations for these high values of C. First of all, it may be that they indicate simply that collaborations of three or more people are common in science.
- Every paper which has three authors clearly contributes a triangle to the numerator of Eq. (2) and hence increases the clustering coefficient. This is, in a sense, a "trivial" form of clustering, although it is by no means socially uninteresting. In fact it turns out that this effect can account for some but not all of the clustering seen in our graphs.
- One can construct a random graph model of a collaboration network which mimics the trivial clustering effect, and the results indicate that only about a half of the clustering which we see is a result of authors collaborating in groups of three or more.
- The rest of the clustering must have a social explanation, and there are some obvious possibilities:

Clustering Coefficients (Explanations of the Structure)

- A scientist may collaborate with two colleagues individually, who may then become acquainted with one another through their common collaborator, and so end up collaborating themselves. This is the usual explanation for transitivity in acquaintance networks.
- Three scientists may all revolve in the same circles—read the same journals, attend the same conferences—and, as a result, independently start up separate collaborations in pairs, and so contribute to the value of C, although only the workings of the community, and not any specific person, is responsible for introducing them.
- As a special case of the previous possibility—and perhaps the most likely case—three scientists may all work at the same institution, and as a result may collaborate with one another in pairs.

Clustering Coefficient of the Biomedical Sciences

- The clustering coefficient of the MEDLINE database is worthy of brief mention, since its value is far smaller than those for the other databases.
- One possible explanation of this comes from the unusual social structure of biomedical research, which, unlike the other sciences, has traditionally been organized into laboratories, each with a "principal investigator" supervising a large number of post docs, students, and technicians working on different projects.
- This organization produces a tree-like hierarchy of collaborative ties with fewer interactions within levels of the tree than between them. A tree has no loops in it, and hence no triangles to contribute to the clustering coefficient. Although the biomedicine hierarchy is certainly not a perfect tree, it may be sufficiently tree-like for the difference to show up in the value of C.
- Another possible explanation comes from the generous tradition of authorship in the biomedical sciences. It is common, example, for a researcher to be made a coauthor of a paper in return for synthesizing reagents used in an experimental procedure.
- Such a researcher will in many cases have a less than average likelihood of developing new collaborations with their collaborators' friends, and therefore of increasing the clustering coefficient.

Distances and Centrality (What is a Geodesic?)

- A fundamental concept in graph theory is the "geodesic," or shortest path of vertices and edges which links two given vertices.
- There may not be a unique geodesic between two vertices: there may be two or more shortest paths, which may or may not share some vertices.

Distances and Centrality (Calculating the Geodesic)

- The geodesic(s) between two vertices i and j can be calculated in time O(m), where m is the number of edges in the graph, using the following algorithm, which is a modified form of the standard breadth-first search

1. Assign vertex j distance zero, to indicate that it is zero steps away from itself, and set $d \leftarrow 0$.
2. For each vertex k whose assigned distance is d, follow each attached edge to the vertex l at its other end and if l has not already been assigned a distance, assign it distance $d+1$. Declare k to be a predecessor of l.
3. If l has already been assigned distance $d+1$, then there is no need to do this again, but k is still declared a predecessor of l.
4. Set $d \leftarrow d+1$.
5. Repeat from step (2) until there are no unassigned sites left.

Now the shortest path (if there is one) from i to j is the path you get by stepping from i to its predecessor, and then to the predecessor of each successive site until j is reached. If a site has two or more predecessors, then there are two or more shortest paths, each of which must be followed separately if we wish to know all shortest paths from i to j.

Watts Barabasi Geodesic

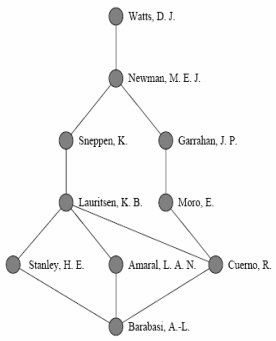


FIG. 4. The geodesics, or shortest paths, in the collaboration network of physicists between Duncan Watts and Laszlo Barabási.

- In Fig. 4 we show the shortest paths of know collaborations between two of the author's colleagues, Duncan Watts (Columbia) and Laszlo Barabasi (Notre Dame), both of whom work on social networks of various kinds.
- It is interesting to note that, although the two scientists in question are well acquainted both personally and with one another's work, the shortest path between them does not run entirely through other collaborations in the field.
- (For example, the connection between the present author and Juan Pedro Garrahan results from our coauthorship of a paper on spin glasses.)
- Although this may at first sight appear odd, it is probably in fact a good sign.
- It indicates that workers in the field come from different scientific "camps," rather than all descending intellectually from a single group or institution.
- This presumably increases the likelihood that those workers will express independent opinions on the open questions of the field, rather than merely spouting slight variations on the same underlying doctrine.

A Database to Facilitate the Collaboration of Computer Scientists

- A database which would allow one conveniently and quickly to extract shortest paths between scientists in this way might have some practical use.
- Kautz has constructed a web-based system which does just this for computer scientists, with the idea that such a system might help to create new professional contacts by providing a "referral chain" of intermediate scientists through whom contact may be established.

Betweeness and Funneling

- A quantity of interest in many social network studies is the "betweenness" of an actor i, which is defined as the total number of shortest paths between pairs of actors which pass through i.
- This quantity is an indicator of who are the most influential people in the network, the ones who control the flow of information between most others.
- The vertices with high betweenness also result in the largest increase in typical distance between others when they are removed.

Betweeness and Funneling (An Efficient Algorithm)

Naively, one might think that betweenness would take time of order $O(mn^2)$ to calculate for all vertices, since there are $O(n^2)$ shortest paths to be considered, each of which takes time $O(m)$ to calculate, and standard network analysis packages such as UCInet [61] indeed use $O(mn^2)$ algorithms at present. However, since breadth-first search algorithms can calculate n shortest paths in time $O(m)$, it seems possible that one might be able to calculate betweenness for all vertices in time $O(mn)$. Here we present a simple new algorithm which performs this calculation. Being enormously faster than the standard packages, it makes possible exhaustive calculation of betweenness on the very large graphs studied here. The algorithm is as follows.

1. The shortest paths from a vertex i to every other vertex are calculated using breadth-first search as described above, taking time $O(m)$.
2. A variable b_k, taking the initial value 1, is assigned to each vertex k.
3. Going through the vertices k in order of their distance from i, starting from the furthest, the value of b_k is added to the corresponding variable on the predecessor vertex of k. If k has more than one predecessor, then b_k is divided equally between them. This means that if there are two shortest paths between a pair of vertices, the vertices along those paths are given betweenness of $\frac{1}{2}$ each.

Betweeness and Funneling (An Efficient Algorithm)

4. When we have gone through all vertices in this fashion, the resulting values of the variables b_k represent the number of geodesic paths to vertex i which run through each vertex on the lattice, with the end-points of each path being counted as part of the path. To calculate the betweenness for all paths, the b_k are added to a running score maintained for each site and the entire calculation is repeated for each of the n possible values of i. The final running scores are precisely the betweenness of each of the n vertices.

Betweeness and Funneling Results

- In column 3 of Table II we show the ten highest betweennesses in the astro-ph, cond-mat, and hep-th subdivisions of the Los Alamos Archive.
- While we leave it to the knowledgeable reader to decide whether the scientists named are indeed pivotal figures in their respective fields, we do notice one interesting feature of the results.
- The betweenness measure gives very clear winners in the competition: the individuals with highest betweenness are well ahead of those with second highest, who are in turn well ahead of those with third highest, and so on.
- This same phenomenon has been noted in other social networks [1].

Funneling

- Strogatz has raised another interesting question about social networks which we can address using our betweenness algorithm: are all of your collaborators equally important for your connection to the rest of the world, or do most paths from others to you pass through just a few of your collaborators?
- One could certainly imagine that the latter might be true. Collaboration with just one or two senior or famous members of one's field could easily establish short paths to a large part of the collaboration network, and all of those short paths would go through those one or two members.
- Strogatz calls this effect "funneling."
- Since our algorithm, as a part of its operation, calculates the vertices through which each geodesic path to a specified actor i passes, it is a trivial modification to calculate also how many of those geodesic paths pass through each of the immediate collaborators of that actor, and hence to use it to look for funneling.

Funneling Results - I

- Our collaboration networks, it turns out, show strong funneling. For most people, their top few collaborators lie on most of the paths between themselves and the rest of the network.
- The rest of their collaborators, no matter how numerous, account for only a small number of paths.
- Consider, for example, the present author.
  - Out of the 44 000 scientists in the giant component of the Los Alamos Archive collaboration network, 31 000 paths from them to me, about 70%, pass through just two of my collaborators, Chris Henley and Juanpe Garrahan.
  - Another 13 000, most of the remainder, pass through the next four collaborators.
  - The remaining five account for a mere 1% of the total.

Funneling Results - II

- To give a more quantitative impression of the funneling effect, we show in Fig. 5 the average fraction of paths that pass through the top 10 collaborators of an author, averaged over all authors in the giant component of the Los Alamos database. The figure shows for example that on average 64% of one's shortest paths to other scientists pass through one's top-ranked collaborator.
- Another 17% pass through the second-ranked one. The top 10 shown in the figure account for 98% of all paths.
- That one's top few acquaintances account for most of one's shortest paths to the rest of the world has been noted before in other contexts. For example, Stanley Milgram, in his famous "small world" experiment [63], noted that most of the paths he found to a particular target person in an acquaintance network went through just one or two acquaintances of the target. He called these acquaintances "sociometric superstars."

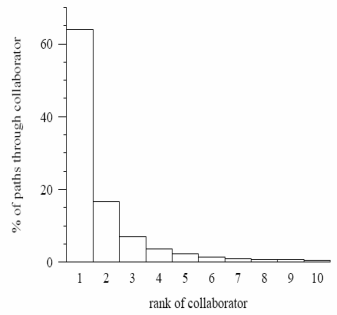


FIG. 5. The average percentage of paths from other scientists to a given scientist which pass through each collaborator of that scientist, ranked in decreasing order. The plot is for the Los Alamos Archive network, although similar results are found for other networks.

Average Distances - I

- Breadth-first search allows to us calculate exhaustively the lengths of the shortest paths from every vertex on a graph to every other (if such a path exists) in time O(mn).
- We have done this for each of the networks studied here and averaged these distances to find the average distance between any pair of (connected) authors in each of the subject fields studied.
- These figures are given in the penultimate row of Table I.
- As the table shows, these figures are all quite small: they vary from 4.0 for SPIRES to 9.7 for NCSTRL, although this last figure may be artificially inflated by the poor coverage of this database discussed in Section IIIE.
- At any rate, all the figures are very small compared to the number of vertices in the corresponding databases.

Small World Effect - I

- This "small world" effect, first described by Milgram [63], is, like the existence of the giant component, probably a good sign for science; it shows that scientific information—discoveries, experimental results, theories—will not have far to travel through the network of scientific acquaintance to reach the ears of those who can benefit by them.
- Even the maximum distances between scientists in these networks, shown in the last row of the table, are not very large, the longest path in any of the networks being just 31 steps long, again in the NCSTRL database, which may have poorer coverage than the others

Small World Effect - II

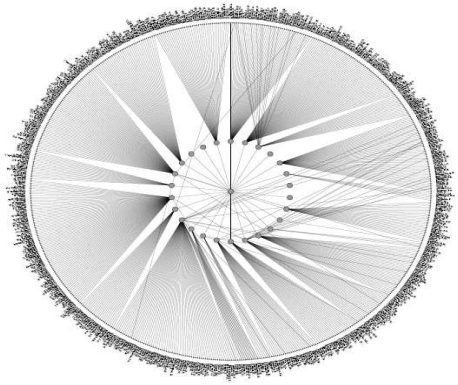


FIG. 6. The point in the center of the figure represents the author of the paper you are reading, the first ring his collaborators, and the second ring their collaborators. Collaborative ties between members of the same ring, of which there are many, have been omitted from the figure for clarity.

Small World Effect - III

- The explanation of the small world effect is simple. Consider Fig. 6, which shows all the collaborators of the present author (in all subjects, not just physics), and all the collaborators of those collaborators—all my first and second neighbors in the collaboration network.
- As the figure shows, I have 26 first neighbors, but 623 second neighbors.

Small World Effect - IV

- The "radius" of the whole network around me is reached when the number of neighbors within that radius equals the number of scientists in the giant component of the network, and if the increase in numbers of neighbors with distance continues at the impressive rate shown in the figure, it will not take many steps to reach this point.
- This simple idea is borne out by theory. In almost all networks, the number of kth-nearest neighbors of a typical vertex increases exponentially with k, and hence the average distance between pairs of vertices ℓ scales logarithmically with n the number of vertices.
- In a standard random graph [56,57], for instance, $\ell = \log n / \log z$, where z is the average degree of a vertex, the average number of collaborators in our terminology.

Small World Effect - V

- In the more general class of random graphs in which the distribution of vertex degrees is arbitrary [58], rather than Poissonian as in the standard case, the equivalent expression is [23]

$$\ell = \frac{\log(n/z_1)}{\log(z_2/z_1)} + 1, \qquad (3)$$

Small World Effect - VI

- where z1 and z2 are the average numbers of first and second neighbors of a vertex.
- It is in fact quite difficult to find a network which does not show logarithmic behavior—such networks are a set of measure zero in the limit of large n.
- Thus the presence of the small world effect is hardly a surprise to anyone familiar with graph theory.
- However, it would be nice to demonstrate explicitly the presence of logarithmic scaling in our networks.
- Figure 7 does this in a crude fashion.
- In this figure we have plotted the measured value of ℓ, as given in Table I, against the value given by Eq. (3) for each of our four databases, along with separate points for nine of the subject-specific subdivisions of the Los Alamos Archive.
- As the figure shows, the correlation between measured and predicted values is quite good. A straight-line fit has R2 = 0.86, rising to R2 = 0.95 if the NCSTRL database, with its incomplete coverage, is excluded (the downward pointing triangle in the figure).

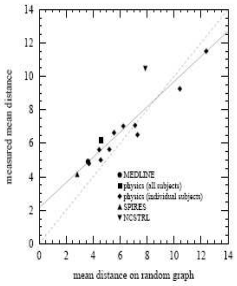


FIG. 7. Average distance between pairs of scientists in the various networks, plotted against average distance on a random graph of the same size and degree distribution. The dotted line shows where the points would fall if measured and predicted results agreed perfectly. The solid line is the best straight-line fit to the data.

Estimating Global Structure from Local Structure

- Turning this observation around, our results also imply that it is possible to make a good prediction of the typical vertex–vertex distance in a network by making only local measurements of the average numbers of neighbors that vertices have.
- If this result ex-tends beyond coauthorship networks to other social networks, it could be of some importance for empirical work, where the ability to calculate global properties of a network by making only local measurements could save large amounts of effort.

Closeness - I

- We can also trivially use our breadth-first search algorithm to calculate the average distance from a single vertex to all other vertices in the giant component.
- This average is essentially the same as the quantity known as "closeness" to social network analysts.
- Like betweenness it is also a measure, in some sense, of the centrality of a vertex—authors with low values of this average will, it is assumed, be the first to learn new information, and information originating with them will reach others quicker than information originating with other sources.
- Average distance is thus a measure of centrality of an actor in terms of their access to information, unlike betweenness which is a measure of an actor's control over information flowing between others.

Closeness - II

- Calculating average distance for many networks returns results which look sensible to the observer.
- Calculations for the network of collaborations between movie actors, for instance, gives small average distances for actors who are famous—ones many of us will have heard of.
- Interestingly, however, performing the same calculation for our scientific collaboration networks does not return sensible results.
- For example, one finds that the people at the top of the list are always experimentalists.
- This, you might think, is not such a bad thing: perhaps the experimentalists are better connected people?
- In a sense, in fact, it turns out that they are. In Fig. 8 we show the average distance from scientists in the Los Alamos Archive to all others in the giant component as a function of their number of collaborators. As the figure shows, there is atrend towards shorter average distance as the number of collaborators becomes large. This trend is clearer still in the inset, where we show the same data averaged over all authors who have the same number of collaborators.
- Since experimentalists work in large groups, it is not surprising to learn that they tend to have shorter average distances to other scientists.

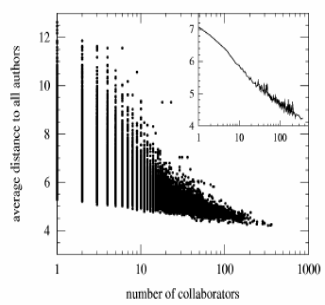


FIG. 8. Scatter plot of the mean distance from each physicist in the giant component of the Los Alamos Archive data to all others as a function of number of collaborators. Inset: the same data averaged vertically over all authors having the same number of collaborators.

Closeness - III

- But this brings up an interesting question, one that we touched upon in Section II: while most pairs of people who have written a paper together will know one another reasonably well, there are exceptions.
- On a high energy physics paper with 1000 coauthors, for instance, it is unlikely that every one of the 499 500 possible acquaintanceships between pairs of those authors will actually be realized.
- Our closeness measure does not take into account the tendency for collaborators in large groups not to know one another, or to know one another less well.
- In the next section we study a more sophisticated form of collaboration network which does do this.

Weighted Collaborative Networks - I

- We can exploit how many papers each pair of scientists has collaborated on during the period of the study, and how many other coauthors they had on each of those papers.
- We can use this information to make an estimate of the strength of collaborative ties.
- Weight collaborative ties inversely according to the number of coauthors as follows.
  - Suppose a scientist collaborates on the writing of a paper which has n authors in total, i.e., he or she has n − 1 coauthors on that paper. Then we assume that he or she is acquainted with each coauthor 1/(n−1) times as well, on average, as if there were only one coauthor.

Weighted Collaborative Networks - II

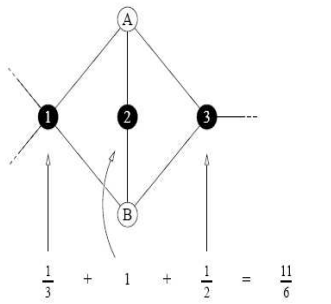


FIG. 9. Authors A and B have coauthored three papers together, labeled 1, 2, and 3, which had respectively four, two, and three authors. The tie between A and B accordingly accrues weight $\frac{1}{3}$, 1, and $\frac{1}{2}$ from the three papers, for a total weight of $\frac{11}{6}$.

- So here is a natural equation to do this (i, j authors, k equation)

$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1}, \qquad (4)$$

- Single author papers are excluded.

Weighted Collaborative Networks - III

Note that the equivalent of vertex degree for our weighted network—i.e., the sum of the weights for each of an individual's collaborations—is now just equal to the number of papers they have coauthored with others:

$$\sum_{j(\neq i)} w_{ij} = \sum_k \sum_{j(\neq i)} \frac{\delta_i^k \delta_j^k}{n_k - 1} = \sum_k \delta_i^k. \qquad (5)$$

Weighted Collaborative Networks - IV

- In Fig. 10 we show as an example collaborations between Gerard Barkema (one of the present author's frequent collaborators), and all of his collaborators in the Los Alamos Archive for the past five years.
- Lines between points represent collaborations, with their thickness proportional to the weights wij of Eq. (4).
- As the figure shows, Barkema has collaborated closely with myself and with Normand Mousseau, and less closely with a number of others.
- Also, two of his collaborators, John Cardy and Gesualdo Delfino have collaborated quite closely with one another.

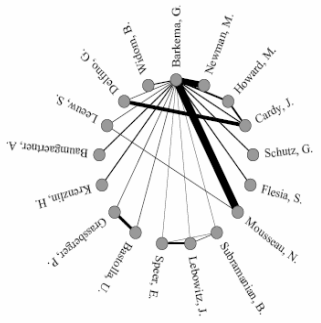


FIG. 10. Gerard Barkema and his collaborators, with lines representing collaborations whose thickness is proportional to our estimate, Eq. (4), of the strength of the corresponding tie.

Geodesic Distances on the Weighted Graph

- In the last column of Table II we show the pairs of collaborators who have the strongest collaborative ties in three subdivisions of the Los Alamos Archive.
- We have used our weighted collaboration graphs to calculate distances between scientists. In this simple calculation we assumed that the distance between authors is just the inverse of the weight of their collaborative tie.
- Thus if one pair of authors know one another twice as well as another pair, the distance between them is half as great.
- Calculating minimum distances between vertices on a weighted graph such as this cannot be done using the breadth-first search algorithm of Section IVA, since the shortest weighted path may not be the shortest in terms of number of steps on the unweighted network.

Calculation of Geodesic Distances on Weighted Graphs Using Dijkstra's Algorithm

1. Distances from vertex i are stored for each vertex and each is labeled "exact," meaning we have calculated that distance exactly, or "estimated," meaning we have made an estimate of the distance, but that estimate may be wrong. We start by assigning an estimated distance of ∞ to all vertices except vertex i to which we assign an estimated distance of zero. (We know the latter to be exactly correct, but for the moment we consider it merely "estimated.")
2. From the set of vertices whose distances from i are currently marked "estimated," choose the one with the lowest estimated distance, and mark this "exact."
3. Calculate the distance from that vertex to each of its immediate neighbors in the network by adding to its distance the length of the edges leading to those neighbors. Any of these distances which is shorter than a current estimated distance for the same vertex supersedes that current value and becomes the new estimated distance for the vertex.
4. Repeat from step (2), until no "estimated" vertices remain.

A naive implementation of this algorithm takes time $O(mn)$ to calculate distances from a single vertex to all others, or $O(mn^2)$ to calculate all pairwise distances. One of the factors of n, however, arises because it takes time $O(n)$ to search through the vertices to find the one with the smallest estimated distance. This operation can be improved by storing the estimated distances in a binary heap (a partially ordered binary tree with its smallest entry at its root). We can find the smallest distance in such a heap in time $O(1)$, and add and remove entries in time $O(\log n)$. This speeds up the operation of the algorithm to $O(mn \log n)$, making the calculation feasible for the large networks studied here.

Average Weighted Distance - I

- It is in theory possible to generalize any of the calculations of Section IV to the weighted collaboration graph using this algorithm and variations on it.
- For example, we can find shortest paths between specified pairs of scientists, as a way of establishing referrals.
- We can calculate the weighted equivalent of betweenness by a simple adaption of our fast algorithm of Section IVB—we use Dijkstra's algorithm to establish the hierarchy of predecessors of vertices and then count paths through vertices exactly as before.
- We can also study the weighted version of the "funneling" effect using the same algorithm.
- Here we carry out just one calculation explicitly to demonstrate the idea; we calculate the weighted version of the distance centrality measure of Section IVC, i.e., the average weighted distance from a vertex to all others.

Average Weighted Distance - II

- Table III we show the winners in this particular popularity contest, along with their numbers of collaborators and papers in the database.
- Many of the scientists who score highly here do indeed appear to be well connected individuals.
- For example, number 1 best connected astrophysicist, Martin Rees, is the Astronomer Royal of Great Britain. (On being informed of this latest honor, Prof. Rees is reported as replying, "I'm certainly relieved not to be the most disconnected astrophysicist"

Average Weighted Distances - III

- What is interesting to note however (apart from nonchalantly checking to see if one has made it into the top 10) is that sheer number of collaborators is no longer a necessary prerequisite for being well-connected in this sense (although some of the scientists listed do have a large number of collaborators).
- The case of D. Youm is particularly startling, since Youm has only three collaborators listed in the database, but nonetheless is fifth best connected high-energy theorist (out of eight thousand), because those three collaborators are themselves very well connected, and because their ties to Youm are very strong. Experimentalists no longer dominate the field, although the well-connected amongst them still score highly.

	rank	name	co-workers	papers
astro-ph:	1	Rees, M. J.	31	36
	2	Miralda-Escude, J.	36	34
	3	Fabian, A. C.	156	112
	4	Waxman, E.	15	30
	5	Colotti, A.	119	45
	6	Narayan, R.	65	58
	7	Loeb, A.	33	64
	8	Reynolds, C. S.	45	38
	9	Hernquist, L.	62	80
	10	Gould, A.	76	79
cond-mat:	1	Fisher, M. P. A.	21	35
	2	Balents, L.	24	29
	3	MacDonald, A. H.	64	70
	4	Sentil, T.	9	13
	5	Das Sarma, S.	51	75
	6	Millis, A. J.	43	37
	7	Ioffe, L. B.	16	27
	8	Sachdev, S.	28	44
	9	Lee, P. A.	24	34
	10	Jungwirth, T.	27	17
hep-th:	1	Cvetic, M.	33	69
	2	Behrndt, K.	22	41
	3	Tseytlin, A. A.	22	65
	4	Bergshoeff, E.	21	39
	5	Youm, D.	3	30
	6	Lu, H.	34	73
	7	Klebanov, I. R.	29	47
	8	Townsend, P. K.	31	54
	9	Pope, C. N.	33	72
	10	Larsen, F.	11	27

TABLE III. The ten best connected individuals in three of the communities studied here, calculated using the weighted distance measure described in the text.

Conclusions

- Interpretation of results
- Computational complexity
- Adaptation to biological problems