Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality

K.BHASKAR RAO[1], TSSN SRINIVAS[2]
ASSOC.PROFESSOR[1], ASST.PROFESSOR[2]
Department of BS & H
BHIMAVARAM INSTITUTE OF ENGINEERING & TECHNOLOGY

ABSTRACT

In the preceding paper [1], we studied collaboration net- works of scientists in which two scientists are considered connected if they have coauthored one or more scientific papers together. As we argued, these networks are for the most part true acquaintance networks, since it is likely that a pair of scientists who have coauthored a paper together are personally acquainted. And since the publication record of scientists is well documented in a variety of publicly avail- able electronic databases, construction of large and relatively complete networks is possible by automated means. These networks provide a promising source of real-world data to fuel the current surge of research interest in social network structure within the physics community.

INTRODUCTION

The networks studied in Ref. [1] were constructed using four publicly available bibliographic databases: Medline, which covers research in biology and medicine; the Los Almost-Print Archive, which covers experimental and theoretical physics; the Stanford Public Information Retrieval Sys- team (SPIRES), which covers experimental and theoretical high-energy physics; and the Networked Computer Science Technical Reference Library (NCSTRL), which covers com- putter science. A broad selection of basic statistics was calculated for these networks, including typical numbers of au- thorns per paper, papers per author, and collaborators per author, as well as distributions of these quantities, existence and size of a giant component, and degree of network clustering. In this second paper, we turn to some more so hesitated, mostly nonlocal, network measures.

DISTANCES AND CENTRALITY

In this section, we look at some measures of network structure having to do with paths between vertices in the network. These measures are aimed at understanding the pat-terns of connection and communication between scientists. In Sec. III we discuss some shortcomings of these measures, and construct some more complex measures that may better reflect true connection patterns.

Shortest paths

A fundamental concept in graph theory is the ‘‘geode- sic,’’ or shortest path of vertices and edges that 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. The geodesic(s) be- tween two vertices i and j can be calculated using the following algorithm, which is a modified form of the standard breadth-first search [2].

(1) Assign vertex j distance zero, to indicate that it is zero steps away from itself, and set d 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 verticals 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 vertex until j is reached. If a vertex 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.

In the standard implementation of this algorithm, a queue (i.e., a first-in/first-out buffer) is maintained of vertices whose distances have been assigned, but whose attached edges have not yet been followed. Using a queue eliminates the need in step 2 above to search through all vertices for those at distance d, and allows the algorithm to run to completion in time $O(m)$, where m is the number of edges in the graph. We note also that the algorithm as we have de- scribed it allows us to calculate the shortest paths from *all* vertices to the target j in a single run, and not just from the single vertex i that we were originally interested in. Thus, we can calculate n shortest paths in time $O(m)$, where n is the

1063-651X/2001/64(1)/016132(7)/$20.00

Betweenness 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 that pass through i [4]. This quantity is an indicator of who the most influential people in the network are, the ones who control the flow of information between most others. The vertices with highest betweenness also result in the largest increase in typical distance between others when they are removed [5]. 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. 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 algorithm that performs this calculation. Being enormously faster than the simple $O(mn^2)$ method, it makes possible exhaustive calculation of betweenness on the very large graphs studied here. The algorithm is as follows.

(1) The shortest paths to a vertex j from 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 j, starting from the farthest, 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 a betweenness of $\frac{1}{2}$ each.

(4) When we have gone through all vertices in this flash- ion, the resulting values of the variables b_k represent the number of geodesic paths to vertex j that 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 vertex and the entire calculation is repeated for each of the n possible values of j. The final running scores are pre- cicely the betweennesses of each of the n vertices.

Using this algorithm, we have been able to calculate be- tweeness exhaustively for all scientists in our networks in reasonable running time. [For example, the calculation for the Los Alamos Archive takes about two hours on a current (*circa* 2000) workstation.] One particularly notable feature of the results is that the betweenness measure gives very clear winners among the scientists in the network: the individual 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 [5].

Stoats [6] has raised an 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 collaborator? 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. Stoats 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 path's pass through each of the immediate collaborators of that actor, and hence to use it to look for funneling.

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 nu- serous, 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 collabo- ration network, 31 000 paths from them to me, about 70%, pass through just two of my collaborators, while another 13 000, most of the remainder, pass through the next four. The remaining five collaborators account for a mere 1% of the total. (These and all other results presented in this paper were calculated using the ''all initials'' versions of our net- works, as described in Ref. [1], except where otherwise noted.)

To give a more quantitative impression of the funneling effect, we show in Fig. 2 the 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 data- base. 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, Milgram, in his famous ''small world'' experiment [7], noted that most of the paths he found to a particular target person in an acquain-tance network went through just one or two acquaintances ofthe target. He called these acquaintances ‘‘sociometric super-stars.’’

A. Average distances

Breadth-first search allows us to calculate exhaustivelythe lengths of the shortest paths from every vertex on a graphto 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 mean 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 because the NCSTRL database appears to have poorer coverage of its subject area than the other databases studied here [1]. At any rate, all the figures are very small compared to the number of vertices in the corresponding databases. This ‘‘small world’’ effect, first described by Milgram [7], is, like the existence of a giant component [1], probably a good sign for science; it shows that scientific information—discoveries, experimental results, theories—will not have far to travel through the net- work of scientific acquaintance to reach the ears of those who can benefit by them. Even the *maximum* distances be- tween scientists in these networks, shown in the last row of Table I, 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.

The explanation of the small world effect is simple. Con-
sider Fig. 3, 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, Ihave 26 first neighbors, but 623 second neighbors. 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 *k*th 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, for instance, $l = \log n/\log z$, where z is the average degree ofa vertex, the average number of collaborators in our terminology [8,9]. In the more general class of random graphs in which the distribution of vertex degrees is arbitrary [10], rather than Poisoning as in the standard case, the equivalent expression is [11]

FIG. 4. Average distance between pairs of scientists in the various networks, plotted against average distance on a random graphof the same size and degree distribution. The dotted line shows where the points would fall if measured and predicted results agreedperfectly. The solid line is the best straight-line fit to the data.

NCSTRL database, with its incomplete coverage, is excluded
(The diamond-shaped symbol in the figure).

Figure 4 needs to be taken with a pinch of salt. Its construction implicitly assumes that the different networks are statistically similar to one another and to random graphs with the same distributions of vertex degree, an assumption that isalmost certainly not correct. In practice, however, the measured value of l seems to follow Eq. (1) quite closely. 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 verticeshave. If this result extends beyond coauthor ship networks to

= log(z 2

other social networks, it could be of some importance for empirical work, where the ability to calculate global proper- ties of a network by making only local measurements could

where z_1 and z_2 are the average numbers of first and second

neighbors of a vertex. It is highly unlikely that a social net- work would not show similar logarithmic behavior— networks that do not are a set of measure zero in the limit of large *n*. The square lattice, for instance, which does not show logarithmic behavior, would be wildly improbable as a topology for a social network. And the introduction of even the smallest amount of randomness into a square lattice or other regular lattice produces logarithmic behavior in the limit of large system size [12,13]. Thus, the small world effect is hardly a surprise to anyone familiar with graph theory. How- ever, it would be nice to demonstrate explicitly the presence of logarithmic scaling in our networks. Figure 4 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. (1) for each of our four databases, along with separate points for ten of the subject-specific subdivisions of the Los Alamos Archive. As the figure shows, the correlation be- tween measured and predicted values is quite good. A straight-line fit has R^2=0.86, rising to R^2=0.95 if the

save large amounts of effort.

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 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, whereas betweenness is a measure of an actor’s control over informa- tion flowing between others.

Calculating average distance for many networks returns results that look sensible to the observer. Calculations for the network of collaborations between movie actors, for in- stance, give small average distances for actors who are famous—ones many of us will have heard of [14]. Interest- ingly, however, performing the same calculation for our sci-

entific collaboration networks does not return sensible re- sults. 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 con-nected people? In a sense, in fact, it turns out that they are. In Fig. 5 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 a trend toward 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 surpris-ing to learn that they tend to have shorter average distances to other scientists.

But this brings up an interesting question: 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 acquaintance-ships between pairs of those authors will actually be realized. Our closeness measure does not take into account the ten- dency for collaborators in large groups not to know one an- other, or to know one another less well, and for this reason the predominance in the closeness rankings of scientists who work in such large groups is probably misleading. In the next section we introduce a more sophisticated form of collabo-

First of all, it is probably the case, as we pointed out at the end of the previous section, that two scientists whose names appear on a paper together with many other coauthors know one another less well on average than two who were the sole authors of a paper. The extreme case that we discussed of a very large collaboration illustrates this point forcefully, but the same idea applies to smaller collaborations too. Even on a paper with four or five authors, the authors probably know one another less well on average than authors from a smaller collaboration. To account for this effect, we weight collabo- rative ties inversely according to the number of coauthors as follows. Suppose a scientist collaborates on the writing of a paper that 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. One can imagine this as meaning that the scientist divides his or her time equally between the n—1 coauthors. This is obviously only a rough approximation: in reality a scientist spends more time with some coauthors than with others. However, in the ab- sence of other data, it is the obvious first approximation to make [16].

Second, authors who have written many papers together will, we assume, know one another better on average than those who have written few papers together. To account for this, we add together the strengths of the ties derived from each of the papers written by a particular pair of individuals [17]. Thus, if 6^{k} is 1 if scientist i was a coauthor of paper k and zero otherwise, then our weight w_{ij} representing the strength of the collaboration (if any) between scientists i and j is $_{i}$

WEIGHTED COLLABORATION NETWORKS

There is more information present in the databases used here than in the simple networks we have constructed from them, which tell us only whether scientists have collaborated or not [15]. In particular, we know on how many papers each pair of scientists has collaborated 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.

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 Sec. II A, since the shortest weighted path may not be the shortest in terms of number of steps on the un-

weighted network. Instead, we use Dijkstra's algorithm [18], which calculates all distances from a given starting vertex i as follows.

(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 low- est 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 that is shorter than a current estimated dis- tance 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 re- main.

A naive implementation of this algorithm takes time $O(mn)$ to calculate distances from a single vertex to all oth- ers, 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 stor- ing the estimated distances in a binary heap (a partially or- dered 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 reduces the time for the evaluation of all pairwise distances to $O(mn \log n)$, making the calculation feasible for the large networks studied here.

It is in theory possible to generalize any of the calcula- tions of Sec. II 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, in $O(m \log n)$ time. We can calcu- late the weighted equivalent of betweenness in $O(mn \log n)$ time by a simple adaptation of our fast algorithm of Sec. II B—we use Dijkstra's algorithm to establish the hierarchy of predecessors of vertices and then count paths through ver- tices exactly as before. We can also study the weighted ver- sion of the ‘‘funneling’’ effect using the same algorithm. For the moment, we have carried out just one calculation explic- itly to demonstrate the idea; we have calculated the weighted version of the closeness centrality measure of Sec. II C, i.e., the average weighted distance from a vertex to all others.The results reveal that, by constrast with the simple close- ness measure, the list of scientists who are well connected in this weighted sense is no longer dominated by experimental- ists, although the well connected among them still score highly; sheer number of collaborators is no longer a good predictor of connectedness. For example, the fifth best con- nected scientist in high-energy theory (out of 8000) is found to have only three collaborators listed in the database, but nonetheless scores highly in our calculation because his ties with those three collaborators are strong and because the collaborators are themselves well connected.

Many of the scientists who score highly in this calculation appear to be well known individuals, at least in the opinion of this author and his colleagues, and are therefore plausibly well connected. We find also that the number of papers writ- ten by scientists who are well connected in this particular sense is universally high. Having coauthored a large number of papers is, as it rightly should be, always a good way of becoming well connected. Whether you write many papers with many different authors, or many with a few, writing many papers will put you in touch with your peers.

II. CONCLUSIONS

We have studied social networks of scientists in which the actors are authors of scientific papers, and a tie between two authors represents coauthorship of one or more papers. The networks studied were based on publication data from four databases in physics, biomedical research, and computer sci- ence. In this second of two papers, we have looked at a variety of nonlocal properties of our networks. We find that typical distances between pairs of authors through the net- works are small—the networks form a ‘‘small world’’ in the sense discussed by Milgram—and scale logarithmically with total number of authors in a network, in reasonable agree- ment with the predictions of random graph models. We have introduced an algorithm for counting the number of shortest paths between vertices on a graph that pass through each other vertex, which is one order of system size faster than previous algorithms, and used this to calculate the so-called ‘‘betweenness’’ measure of centrality on our graphs. We also show that for most authors the bulk of the paths between them and other scientists in the network go through just one or two of their collaborators, an effect that Strogatz has dubbed ‘‘funneling.’’

We have suggested a measure of the strength of collabo- rative ties which takes account of the number of papers a given pair of scientists have written together, as well as the

number of other coauthors with whom they wrote those pa- pers. Using this measure we have added weightings to our collaboration networks and used the resulting networks to find which scientists have the shortest average distance to others. Generalization of the betweenness and funneling cal- culations to these weighted networks is also straightforward.The calculations presented in this paper and the preceding one inevitably represent only a small part of the investiga- tions that could be conducted using large

network data sets such as these. Indeed, one of the primary intents of this paper is simply to alert other researchers to the presence of a valu- able source of network data in bibliographic databases. We hope, given the high current level of interest in network phe- nomena, that others will find many further uses for these
data.

The author recently learned of a report by Brandes [19] in which an algorithm for calculating betweenness similar to ours is described. The author is grateful to Rick Grannis for bringing this to his attention.

ACKNOWLEDGMENTS

The author would particularly like to thank Paul Ginsparg for his invaluable help in obtaining the data used for this study. The data were generously made available by Oleg Khovayko, David Lipman, and Grigoriy Starchenko (Med- line), Paul Ginsparg and Geoffrey West (Los Alamos e-Print Archive), Heath O'Connell (SPIRES), and Carl Lagoze (NCSTRL). The Los Alamos e-Print Archive is funded by the NSF under Grant No. PHY-9413208. NCSTRL is funded through the DARPA/CNRI test suites program under DARPA Grant No. N66001-98-1-8908. The author would also like to thank Steve Strogatz for suggesting the ''funnel- ing effect'' calculation of Sec. II B, and Dave Alderson, László Barabási, Lin Freeman, Paul Ginsparg, Rick Grannis, Jon Kleinberg, Vito Latora, Sid Redner, Ronald Rousseau, Steve Strogatz, Duncan Watts, and Doug White for many useful comments and suggestions. This work was funded in part by the National Science Foundation and Intel Corpora- tion.

REFERNCE

[1] M.E.J. Newman, preceding paper, Phys. Rev. E **64**, 016131 (2001).
[2] R. Sedgewick, *Algorithms* (Addison-Wesley, Reading, MA, 1988).
[3] H. Kautz, B. Selman, and M. Shah, Commun. ACM **40**, 63 (1997).
[4] L.C. Freeman, Sociometry **40**, 35 (1977).
[5] S. Wasserman and K. Faust, *Social Network Analysis* (Cam-bridge University Press, Cambridge, 1994).
[6] S.H. Strogatz (private communication).
[7] S. Milgram, Psychol. Today **2**, 60 (1967).
[8] P. Erdös and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[9] B. Bollobás, *Random Graphs* (Academic Press, New York, 1985).
[10] M. Molloy and B. Reed, Random Struct. and Algorithms **6**, 161 (1995); Combinatorics, Prob. Comput. **7**, 295 (1998).
[11] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, e-print cond-mat/0007235.
[12] M. Barthélémy and L.A.N. Amaral, Phys. Rev. Lett. **82**, 3180 (1999).
[13] M.E.J. Newman and D.J. Watts, Phys. Rev. E **60**, 7332 (1999).
[14] See www.cs.virginia.edu/oracle/center_list.html.
[15] A complete description of a collaboration network, or indeed any affiliation network, requires us to construct a bipartite graph or hypergraph of actors and the groups to which they belong [5,11]. For our present purposes, however, such de- tailed representations are not necessary.
[16] One can imagine using a measure of connectedness that weights authors more or less heavily depending on the order in which their names appear on a publication. We have not adopted this approach here, however, since it will probably discriminate against those authors with names falling toward the end of the alphabet, who tend to find themselves at the ends of purely alphabetical author lists.

[17] In the study of affiliation networks it is standard to weight ties by the number of common groups to which two actors belong [5], which would be equivalent in our case to taking frequency of collaboration into account but not number of co-workers. In the physics literature this method has been used, for example, to study the network of movie actors [M. Marchiori and V. Latora, Physica A **285**, 539 (2000)].

[18] R.K. Ahuja, T.L. Magnanti, and J.H. Orlin, *Network Flows: Theory, Algorithms, and Applications* (Prentice-Hall, Engle- wood Cliffs, NJ, 1993).

[19] U. Brandes, University of Konstanz report, 2000 (unpub- lished).

