Analysis of SIGMOD's Co-Authorship Graph

Mario A. Nascimento
Dept. of Computing Science
University of Alberta
Canada
mn@cs.ualberta.ca

Jörg Sander
Dept. of Computing Science
University of Alberta
Canada
joerg@cs.ualberta.ca

Jeffrey Pound
Dept. of Computing Science
University of Alberta
Canada
pound@cs.ualberta.ca

ABSTRACT

In this paper we investigate the co-authorship graph obtained from all papers published at SIGMOD between 1975 and 2002. We find some interesting facts, for instance, the identity of the authors who, on average, are "closest" to all other authors at a given time. We also show that SIGMOD's co-authorship graph is yet another example of a small world—a graph topology which has received a lot of attention recently. A companion web site for this paper can be found at `http://db.cs.ualberta.ca/coauthorship`.

1. INTRODUCTION

In the Fall'2002 edition of SIGIR Forum, Smeaton et al [1] analysed the past 25 years of SIGIR publications. Among the analyses done, a particularly interesting one was exploring the graph induced by the co-authorship relations. Given that DBLP[1] is a reliable source of similar data for a number of conferences and journals, and it also makes this data freely available in XML format, it seemed natural to do a similar analysis for other conferences.

First, we extracted all authorship data for entries related to publications—research and industrial full papers, demos, tutorials and panels (chairs)—at SIGMOD conferences between 1975 and 2002 from DBLP's XML data file. There are several statistics one can obtain easily from this data set. For instance, the number of authors per paper varied between 1 and 18. However, 29% of all papers are authored by only one author and about 90% of the papers are co-authored by 4 authors or less. A related question is: how many SIGMOD authors had one paper at SIGMOD, how many had two papers, and so on? Table 1 shows these numbers as well as the accumulated percentage of authors that had less or equal to a certain number of papers at SIGMOD. About 70% of the authors have just one paper at SIGMOD. The vast majority, i.e., over 90% of the authors have 3 or less papers, but single individuals may have significantly larger numbers of papers at SIGMOD, up to 32.

More information can be obtained if we use this data to derive a co-authorship graph, as discussed in the following.

2. IT IS A SMALL WORLD AFTER ALL

Based on DBLP's dataset, we built several co-authorship graphs $G_y(V, E)$, where V is the set of nodes, corresponding to all SIGMOD authors up to year y, and E is the set of edges connecting nodes from V, representing all co-authorship relationships up to year y.

[1] `http://dblp.uni-trier.de/`

Table 1: Number of papers per author at SIGMOD

# Papers	# Authors	% of Authors with ≤ #Papers
32	1	100.00%
27	1	99.96%
26	1	99.92%
24	1	99.87%
21	1	99.83%
20	1	99.79%
18	1	99.75%
17	1	99.71%
16	3	99.67%
15	7	99.54%
14	3	99.25%
13	4	99.12%
12	3	98.95%
11	6	98.83%
10	4	98.58%
9	9	98.41%
8	14	98.04%
7	20	97.45%
6	27	96.62%
5	44	95.49%
4	62	93.65%
3	135	91.06%
2	362	85.42%
1	1683	70.30%

As one would expect, in no year did the co-authorship graph consist of only one single connected component. Up until 1998 the largest connected component had less than half of the total number of nodes. Since then, this ratio has been increasing smoothly. In 2002 the largest connected component had 1413 out of 2394 nodes, i.e., about 60% of SIGMOD authors are directly or indirectly connected to each other. The size of the second largest component presents a different behavior. It has never been larger than 47 (in 1989) and, as of 2002, it has only 11 nodes, i.e., less than 0.5% of all authors.

Since some of the properties discussed next assume the graph to be a single connected component, from now on we consider only the largest connected component of the current co-authorship graph, $G_{2002}(V, E)$, unless noted otherwise.

According to Watts [2], the *characteristic path length* of a graph G is defined as the average shortest path length between every pair of vertices in G. The *clustering coefficient* measures how well the direct neighbors of a vertex

are connected among themselves. For a given node v let $G'(V', E')$ be the subgraph where V' is the set of direct neighbors of v and E' is the set of edges from E between the nodes in V'. Then, the clustering coefficient of v is defined as $\frac{|E'|}{(|V'|\times(|V'|-1))/2}$, measuring the number of edges between the direct neighbors of v as a fraction of all edges that could theoretically exist between them. The average clustering coefficient over all nodes in G is the clustering coefficient of G.

Regular graphs, e.g., regular lattices, are characterized by a high clustering degree, i.e., neighbors of nodes are very well connected among themselves, but they also have a high characteristic path length, meaning that distance between nodes (if they are not direct neighbors) is typically large.

A graph $G(V, E)$ is said to be a *small world graph* if it is characterized by the following two conditions: (1) It has a much higher clustering coefficient than similarly sized random graphs. The clustering coefficient is in fact much closer to the clustering coefficient of a regular graph where the neighbors of nodes are very well connected among each other. (2) It has only a slightly larger characteristic path length than similarly sized random graphs. This means that like in a random graph (and in contrast to regular graphs) the distance between nodes, even if they are not direct neighbors, is typically short, requiring only a few steps in the shortest path between them.

Small world graphs have received a lot of attention recently because they are abundant in nature and man-made systems and the distinctive combination of high clustering degree and small characteristic path length often has important dynamical consequences for the network [2].

We compared the SIGMOD co-authorship graph to random graphs built with the same number of authors (1413) and the same number of edges (4252). The random graphs yielded, on average, a clustering coefficient of 0.004 and an average characteristic path length of 4.24, whereas SIGMOD's graph (as of 2002) yields a clustering coefficient of 0.69 and a characteristic path length of 5.65. Hence, the current SIGMOD co-authorship graph is a small world graph.

The typical behavior one can expect from SIGMOD's co-authorship graph is that, after the establishment of a significant largest component, the characteristic path length will vary slightly with an overall tendency to grow only very smoothly over time. Figure 1 shows the actual evolution of the characteristic path length for SIGMOD's graph. Between 1975 and 1979, the characteristic path length of the largest connected component is very small (between 2 and 3), but there is also not a significant largest connected component, e.g., in 1979, the component sizes ordered by the number of authors was 16, 11, 9, 8, 6, 5, 4, 3, and 2 (for component sizes smaller than 6, several components of that size existed). Then, in 1980 the characteristic path length almost doubles. This is the combined effect of two events that happened in that year. First, the number of authors in the graph increased by 32%, the largest increase in the history of SIGMOD (the average yearly increase, excluding 1980, is only around 9%). Second, the size of the largest connected component grows by a factor of 3, from 16 to 65 (the average yearly growth rate, excluding 1980, is only 23%). This increase in size is not only due to the new authors (not all of them were connected to the largest connected component), but also a consequence of several smaller independent components being connected to the largest connected component in 1980. This effect can be expected to happen. Eventually, authors from different components who had not previously co-authored a paper may write one together and connect the two components, thus increasing the characteristic path length. The opposite effect, i.e., a decrease in the characteristic path length, can be observed in some cases, e.g., if two authors who were not directly connected, write a paper together. An extreme case of this event can be seen in 1986 when L.A. Rowe co-authored a paper with M. Stonebraker. At that point both where already fairly well connected to a lot of other authors, but not directly. Their paper built a bridge or shortcut between two well connected nodes, bringing everyone in the largest connected component closer together. A case of the former effect (although less pronounced than in 1980) is the increase of the characteristic path length in 1989. The reason is a paper co-authored by G.M. Lohman, J.C. Freytag, L.M. Haas, and H. Pirahesh which appeared in 1989. This paper connected two of the smaller components of size 19 and 8 to the largest connected component of size 171 at that time, increasing the characteristic path length significantly, since it introduced paths between all nodes in the formerly disconnected components via a single "bridge".

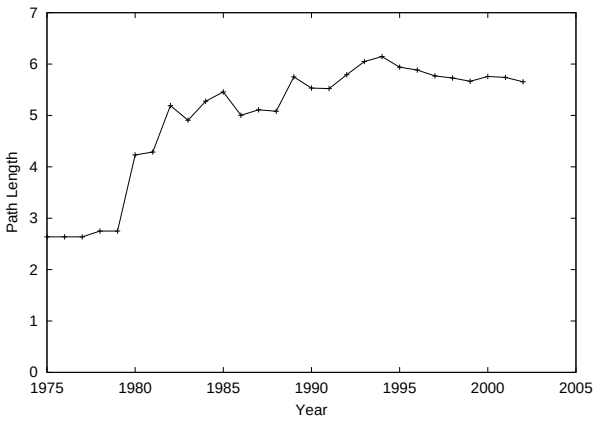

Figure 1: Evolution of the characteristic path lengths.

The graph's diameter, i.e., the *maximum* distance between two authors in the largest connected component of SIGMOD's co-authorship graph follows a similar pattern as the characteristic path length, as the network grows over time, only its values are larger. In the beginning, the diameter of the co-authorship graph was 6 until 1979, when it jumped to 12 over the next three years as the size of largest connected component of the graph grew significantly. Since 1996 the diameter has been constant at a value of 15.

3. VOYAGE TO THE CENTER OF THE CO-AUTHORSHIP GRAPH

One property of a graph that can lead to a small world is the existence of so-called hubs, i.e., nodes that have an extremely high outdegree (direct co-authors) compared to other nodes, and that act as shortcuts to connect nodes to each other. Figure 2 indicates the existence of such hubs, showing that a relatively small number of authors have a very large outdegree while the majority of the authors have

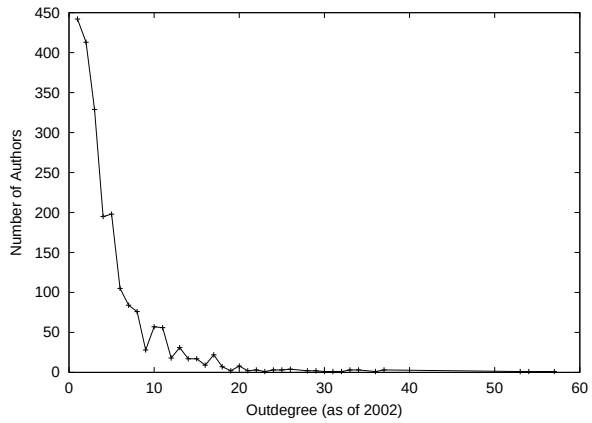

Figure 2: Outdegree distribution of SIGMOD's largest connected component.

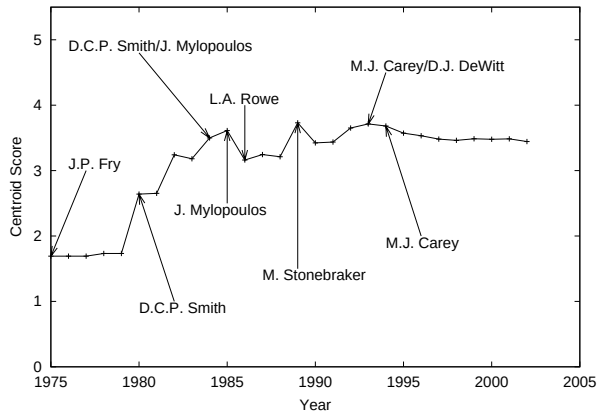

Figure 3: Minimum average path length over time.

much smaller outdegrees. In graphs where hubs exist, it makes sense to measure degrees of centrality of nodes in order to distinguish how close a node is on average to all other nodes. We define the *centrality score* of a node (author) as the average of the shortest path lengths between that node and all others in the largest connected component. If we sort all nodes by the ascending order of their centrality score, we obtain, the authors that, on average, are closer to all other authors in the co-authorship graph at the top of that list. Figure 3 shows the evolution of the minimum centrality score for all nodes over the given years. Author names are displayed in the year they first hold the minimum score, and the names are not repeated until another author has a new minimum centrality score. Qualitatively, the curve is similar to the one in Figure 1 which indicates the strong correlation between the characteristic path length and minimum centrality score. Note that other authors can have centrality scores very close to the minimum one. For instance, Table 2 shows the ten smallest centrality scores (and their holders) as of 2002, and the largest centrality score, out of 1413 authors, is less than three times greater than the smallest one.

Other statistics reported in the SIGIR Forum paper, and which can now be trivially done with SIGMOD's data were, for instance, who are the authors with largest numbers of

Table 2: Ten smallest centrality scores as of 2002.

Author	Avg. Minimum Path Length
Michael J. Carey	3.44
Jeffrey F. Naughton	3.53
Rakesh Agrawal	3.55
Hamid Pirahesh	3.68
David J. DeWitt	3.68
Bruce G. Lindsay	3.72
Johannes Gehrke	3.76
Michael Stonebraker	3.77
Divesh Srivastava	3.80
Jennifer Widom	3.81

papers and co-authors. In SIGMOD's case there is a strong correlation between the number of co-authors and the number of papers a well-connected author has. Indeed, seven of the authors in Table 2 are among the top ten authors with most SIGMOD papers, and six of those are also among the top ten authors with most co-authors.

4. CONCLUSION

Related analyses have been done elsewhere. A well known example is the Erdös Number Project[2], where someone's Erdös number is the minimum path length between that person and P. Erdös, a famous mathematician, also in a co-authorship graph. Another example is the Oracle of Bacon[3], where in addition to the Bacon Number, which is the equivalent to the Erdös Number but with respect to actor Kevin Bacon in the "co-starring in a movie" relationship graph, the centrality score of every actor is also computed.

The findings discussed above are a representative sample of interesting facts one can draw about the co-authorship linkage within SIGMOD. As DBLP's date becomes more complete, it will allow one to explore this type of data even further, e.g., analogous to the notion of centrality in the co-authorship graph, one can compute a notion similar to centrality scores in the cross-referencing graph.

We have also setup a web site[4] with a summary of similar findings for a few other conferences, in addition to SIGMOD, namely: PODS, VLDB and ICDE—the user can choose to use data from either one or from all conferences combined.

Acknowledgments

We thank Tamer Özsu and Rick Snodgrass for the enthusiastic support and interesting discussions. Michael Ley's effort in maintaining DBLP has been instrumental for the analysis presented herein. This work was partially funded by NSERC Canada.

5. REFERENCES

[1] A.F. Smeaton et al. Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century ? *SIGIR Forum*, 36(2):39–43, 2002.

[2] D. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, 1999.

[2] `http://www.oakland.edu/~grossman/erdoshp.html`

[3] `http://www.oracleofbacon.org/`

[4] `http://db.cs.ualberta.ca/coauthorship`