On Six Degrees of Separation in DBLP-DB and More

Ergin Elmacioglu
Department of Computer Science and Engineering
The Pennsylvania State University, PA 16802
elmaciog@cse.psu.edu

Dongwon Lee
School of Information Sciences and Technology
The Pennsylvania State University, PA 16802
dongwon@psu.edu

ABSTRACT

An extensive bibliometric study on the db community using the collaboration network constructed from DBLP data is presented. Among many, we have found that (1) the average distance of all db scholars in the network has been stabilized to about 6 for the last 15 years, coinciding with the so-called six degrees of separation phenomenon; (2) In sync with Lotka's law on the frequency of publications, the db community also shows that a few number of scholars publish a large number of papers, while the majority of authors publish a small number of papers (i.e., following the power-law with exponent about -2); and (3) with the increasing demand to publish more, scholars collaborate more often than before (i.e., 3.93 collaborators per scholar and with steadily increasing clustering coefficients).

1. INTRODUCTION

Social network analysis is an active research field in the social sciences where researchers try to understand social influence and groupings of a set of people or groups. Its origin is in general believed to be due to S. Milgram [7] in 1967 who identified the so-called "*six degrees of separation"* phenomenon based on an experiment – any two people in the United States are connected through about 6 intermediate acquaintances, implying we live in a rather small-world. Since then, sociologists and psychologists have found evidence for a wide range of small-world phenomena arising in other social and physical networks (e.g., power grids, airline time tables, food chain, World-Wide Web, Erdos number). Inspired by some of the recent attempts to apply social network analysis to our own db community [8, 11], in this paper, we analyze the collaboration network made of by database researchers, and see if there exists any interesting patterns underlying the db community and their publication behavior.

2. SETUP

Since DBLP [3] is a high-quality citation digital library that has a near complete coverage on database literature, we chose to use DBLP as the data set for our analysis of the db community. In particular, we examined citation data in DBLP from 1968 to 2003, and hand-picked publication venues (19 journals and 81 conferences, symposiums and workshops) that we believed to be closely-related to the db community (shown in Table 1). Note that we intentionally excluded venues related to Information Retrieval or Digital Library, but included ones related to Data Mining. We also did not include venues that have some database papers as well as papers from many other fields (e.g., J. ACM, Comm. ACM, and WWW). Hereafter, we will refer to this collection as **DBLP-DB**. DBLP-DB contained 32,689 authors and 38,773 papers.

Table1: The list of publication venues in DBLP-DB from 1968 to 2003.

Conferences/Symposiums/Workshops (81)
ADB, ADBIS, ADBT, ADC, ARTDB, Berkeley Workshop, BNCOD, CDB, CIDR, CIKM, CISM, CISMOD, COMAD, COODBSE, CoopIS, DAISD, DANTE, DASFAA, DaWaK, DBPL, DBSEC, DDB, DDW, DEXA, DIWeB, DMDW, DMKD, DNIS, DOLAP, DOOD, DPDS, DS, EDBT, EDS, EFIS/EFDBS, ER, EWDW, FODO, FoIKS, FQAS, Future Databases, GIS, HPTS, IADT, ICDE, ICDM, ICDT, ICOD, IDA, IDC(W), IDEAL, IDEAS, IDS, IGIS, IWDM, IW-MMDBMS, JCDKB, KDD, KR, KRDB, LID, MDA/MDM, MFDBS, MLDM, MSS, NLDB, OODBS, OOIS, PAKDD, PKDD, PODS, RIDE, RIDS, RTDB, SBBD, SDM-SIAM, Semantics in Databases, SIGMOD, SSD, SSDBM, SWDB, TDB, TSDM, UIDIS, VDB, VLDB, WebDB, WIDM, WISE, XP, XSym
Journals (19)
ACM TODS, ACM TOIS, DKE, Data Base, DMKD, DPD, IEEE Data Eng. Bulletin, IEEE TKDE, Info. Processing and Management, Info. Processing Letters, Info. Sciences, Info. Systems, J. of Cooperative Info. Systems, J. of Database Management, JIIS, KAIS, SIGKDD Explorations, SIGMOD Record, VLDB J.

The *collaboration network* (or graph) consists of nodes of authors and edges connecting any two authors if they co-authored one or more papers. Note that DBLP itself does not have a notion of "unique key" such as DOI (Digital Object Identifier). Instead, DBLP depends on the name of authors to distinguish them. Therefore, the classical *name authority control problem* may arise (i.e., same author with various spellings or different authors with the same spelling). We could minimize this problem as Newman [10] did by conducting two experiments – one with full names (``John Doe") and the other with the first initial of the first name followed by the last name (``J. Doe") – and use these as the upper and lower bounds. However since it is known that their effect on the quality of citation analysis

is negligible [10], we did not do any special pre-processing to handle such cases. For the visualization of our network analysis, we used Pajek [2] and NetDraw [9].

3. STATISTICS ABOUT AUTHORS

First, we do various statistical analysis related to authors of papers. Figure 1 shows the number of "new authors" (ones who publish a paper in DBLP-DB for the first time) as a function of year. As shown, the db community steadily grows each year by the addition of new authors (at the 10% rate after 1985). In 2003 alone, there are more than 3,000 new authors who joined the db community, some of whom are novice graduate students or veteran scholars from similar fields.

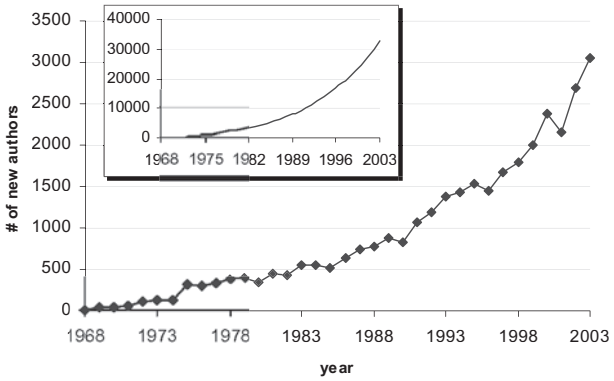

Figure 1. Number of new authors who joined the network each year. Inset: cumulative number of authors up to the year indicated.

Next, we examine how active the db community is. The "active authors" are those who publish at least one paper in a given year. Figure 2 illustrates the number of active authors each year. For instance, in 2003, there are only about 6,000 active authors (out of 32,689 authors). Interestingly, almost half of the active authors in any given year are "new authors". Moreover, new authors are steadily contributing to about 60% of papers each year (since 1982).

Those could be new graduate students entering into the community for the first time by collaborating with their advisors. The remaining 40% of the publications is contributed purely by the existing authors or co-authors, which increases the density of the collaboration network.

Figure 3 illustrates the average number of papers per author for a given year (i.e., # of papers / # of authors). After 1980, the value starts to stabilize around 0.3 paper per author ratio, implying that the productivity rate of the db community as a whole remains intact over time. This makes sense since only small fractions of the community (about 18%-20%) are active each year and they can publish only a limited number of papers.

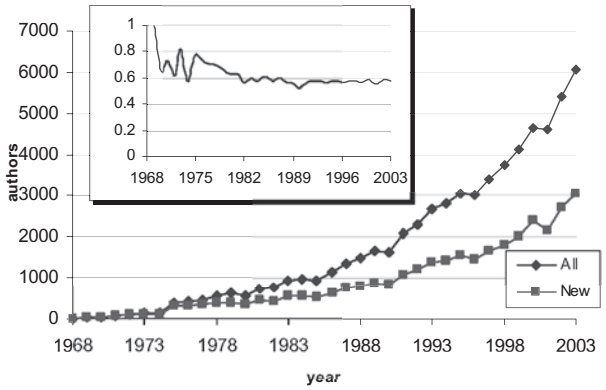

Figure 2. Number active authors each year. Inset: percentage of the papers published by new authors each year

Lotka's Law describes the frequency of publications by authors by "the number of authors making *n* contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 percent" [6]. He showed that such a distribution follows a power law with an exponent approximately -2. Figure 4 shows the distribution of numbers of papers per author on log-log scales for our database.

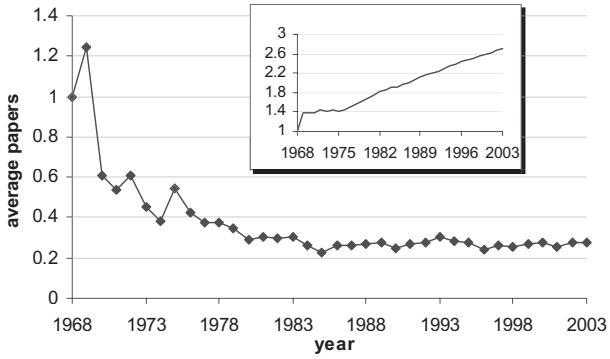

Figure 3. Average number papers per author each year. Inset: cumulative number of papers/author up the year indicated.

Consistent with Lotka's Law, 20849 authors (64%) have only one paper whereas a small number of authors publish a large number of papers (the fat tail on the right hand side indicates this). In fact, in DBLP-DB, there are only 18 authors who published more than 100 papers. Furthermore, the exponent of the graph is -2.15 which is very close to that found by Lotka. Top-10 authors with the highest number of publications in DBLP-DB are shown in Table 4.

Now, we examine the number of collaborators of authors. Figure 5 illustrates the results of this measure. The average number of collaborators per author (from 1968 to 2003) is 3.93 and tends to increase steadily. Compared to

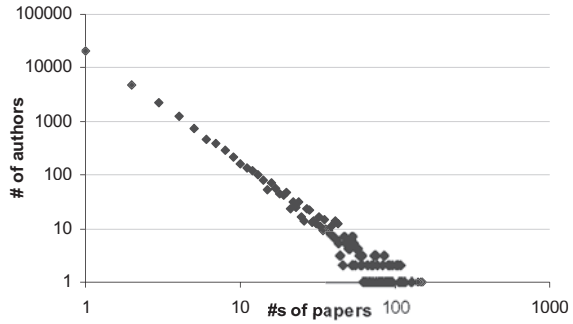

Figure 4. Distribution of numbers of papers per author, as of 2003.

other scientific communities that involves large-scale experimentation (e.g., high-energy physics), this average number of collaborators is rather small.

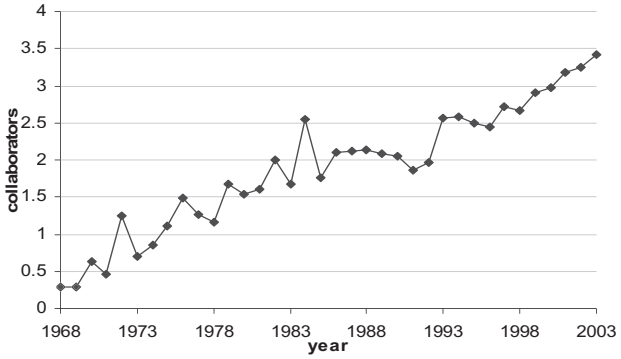

Figure 5. Average number of collaborators per author each year

The steady increase of the average number of collaborators can be hypothesized as follows: (1) the so-called "Publish or Perish" pressure drives scholars to seek more effective ways to increase the number of publications such as collaborative research; and (2) the rapid development and deployment of new communication mediums (email, messenger, web board, or web camera) makes remote collaborations much easier than before.

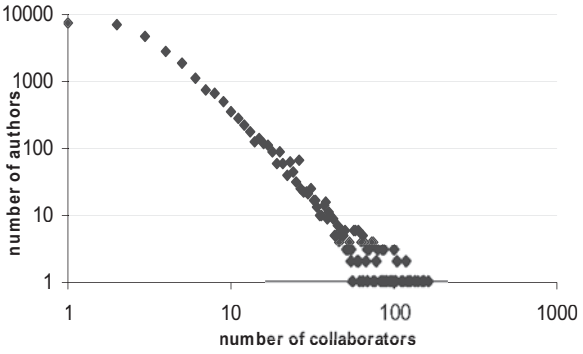

Figure 6. Distribution of number of collaborators per author, as of 2003.

The distribution of number of collaborators per author is shown in Figure 6. It also exhibits the power-law tail with exponent -2.3. The second column of Table 4 shows the authors with the largest number of collaborators. Many of these authors are ranked high in the centrality measures described in section 8.

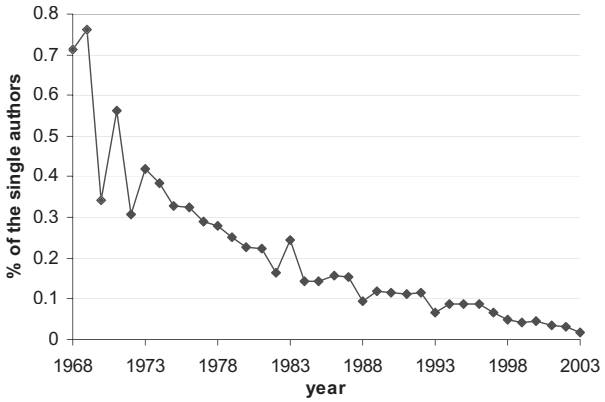

Figure 7. Percentage of the active single authors per year

Finally, we analyzed if there are authors with no collaborators in DBLP-DB. There are 3,073 such authors which constitute 9.4% of the db community. However, 84% of all these single authors have only one paper. There are also a few scholars who have written more than 10 papers just by themselves without any collaboration. One particular author, "*Levent V. Orman*", has written 14 papers alone, shown in Table 2. However, due to the pressure for increased productivity, such single authors are diminishing. Figure 7 shows the percentage of single authors to active authors each year, clearly decreasing over time, and more interestingly, it exhibits the symmetric pattern from the increasing pattern of the number of collaborators in Figure5.

Table 2: Publications of the most productive author with no collaborators.

Year	Title
1982	A familial model of data for a multilevel schema framework.
1984	A Multilevel Design Arct. for decision support systems
1984	Nested set languages for functional databases.
1985	Functions in Information Systems.
1985	Design criteria for functional data bases.
1986	Functional data model design.
1986	Redundancy in functional databases.
1991	Complexity of database languages.
1991	A visual data model.
1993	Knowledge Management by Example.
1996	Queries = Examples + Counterexamples.
1998	Differential Relational Calculus for Integrity Maintenance.
1998	Storage and Retrieval of Database Constraints.
2001	Transaction Repair for Integrity Enforcement.

4. STATISTICS ABOUT PAPERS

In 2003 alone, there are more than 3,000 db papers published (Figure 8), and at the end of 2004, there will be roughly 10% more db papers published. This is largely due to the increased number of authors. Since the average number of papers per author is fixed to 0.3 per year (Figure 3), the number of db papers is approximately in sync with the number of db authors, especially active ones (Figure 2).

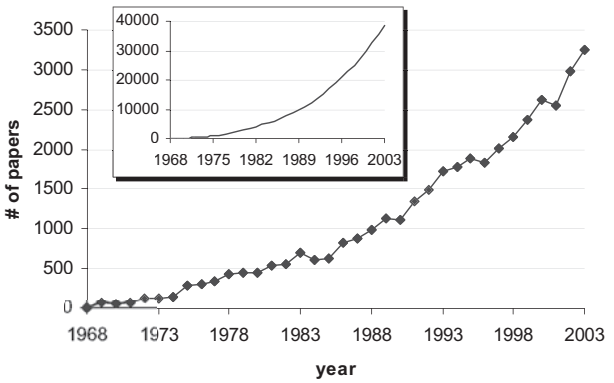

Figure 8. Number of papers per year. Inset: cumulative number of papers published up to the year indicated.

The average number of authors per paper in the db community tends to increase each year, yielding almost 2.8 co-authors per paper as of 2003 (Figure 9). Although there are a significant number of papers with only a single author, there are more papers written by two authors (13557 papers). This can be seen in the distribution of this measure in Figure 10, which has a power law tail with an exponent -3.68. The largest number of authors on a single paper is 27. The figure clearly shows that there is an increasing tendency for collaboration among authors which also causes papers to have more co-authors.

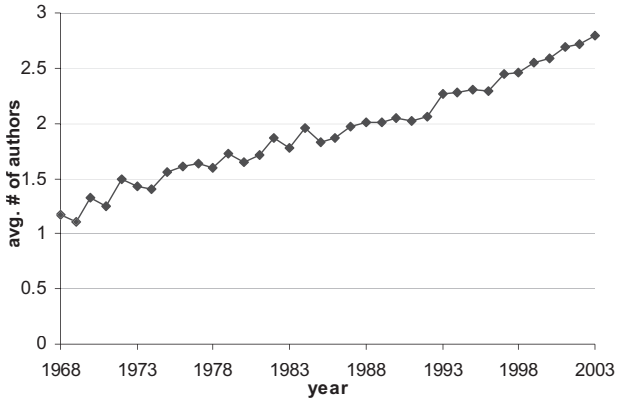

Figure 9. Average number of authors per paper each year

Next, we looked at how publication venues in DBLP-DB are inter-related to each other using co-authorship information. By examining the pattern where the db scholars publish their papers, one can see, for instance,

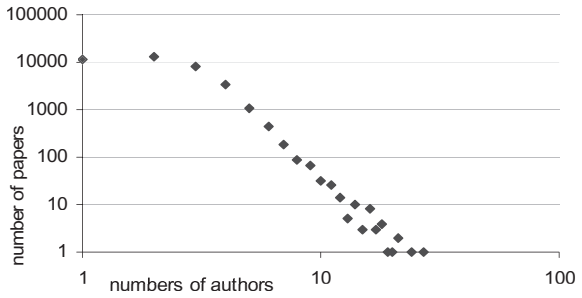

Figure 10. Distribution of number of authors per paper (2003).

which publications venues have a similar theme or taste. Figure 11 is a graph where (1) a node is a publication venue in DBLP-DB, the size of which is proportional to the number of papers in it, and (2) an edge between venues X and Y reflects the Jaccard distance, $|A \cap B| / |A \cup B|$, where A and B are author sets of venues X and Y. The higher the Jaccard distance is (i.e., more authors are common between venues), the thicker the edge becomes. Table 3 lists top-10 pairs of database publication outlets.

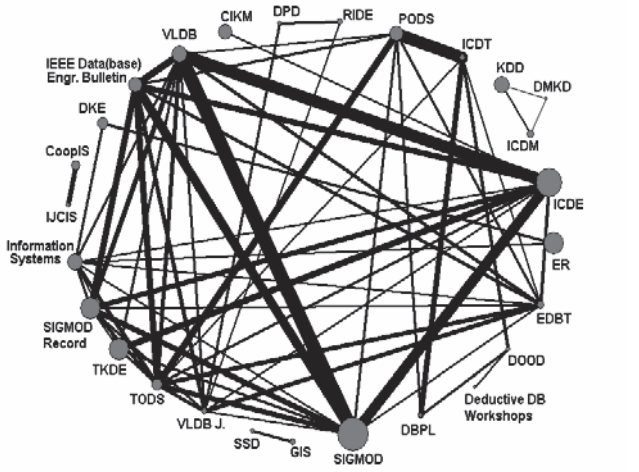

Figure 11: Venue relations (only venues with at least 100 papers and edges with at least 0.1 Jaccard distance are shown. The size of a node is proportional to the number of papers in the venue, while the thickness of edge is proportional to the overlap of authors between venues).

Table 3: Top-10 pairs of venues with the highest Jaccard distances

Similar venue pair	Distance
SIGMOD - VLDB	0.2229
ICDT - PODS	0.1971
ICDE - VLDB	0.1948
ICDE - SIGMOD	0.1817
SIGMOD - IEEE Data Eng. Bulletin	0.1736
VLDB - IEEE Data Eng. Bulletin	0.1559
PODS - TODS	0.1557
SIGMOD Rec. - IEEE Data Eng. Bulletin	0.1502
ICDE - TKDE	0.1498
TODS - IEEE Data Eng. Bulletin	0.1441

Table 4: The top-10 db authors with the highest number of papers, number of co-authors, and clustering coefficients (Clustering coeff. I with 60 as the threshold number of co-authors and clustering coeff. II with 60 as the threshold number of papers). For the number of papers and co-authors columns, values in the parenthesis are for the whole DBLP data set.

Number of papers		Number of co-authors		Clustering coeff. (I)		Clustering coeff. (II)	
151 (293)	H. Garcia-Molina	163 (192)	M. Stonebraker	0.1627	J. F. Roddick (60)	0.2019	A. Segev (68)
147 (201)	J. Han	152 (181)	M.J. Carey	0.1534	M. L. Brodie (65)	0.1498	D. B. Lomet (102)
146 (207)	M. Stonebraker	150 (206)	D. Maier	0.1441	R. Rastogi (60)	0.1441	R. Rastogi (74)
142 (326)	P.S. Yu	139 (200)	H. Garcia-Molina	0.1384	T. Milo (60)	0.1275	Y. Manolopoulos (62)
128 (293)	E. Bertino	134 (165)	D.J. DeWitt	0.1378	B. G. Lindsay (85)	0.1240	J. Widom (90)
115 (150)	R. Agrawal	125 (155)	J. Han	0.1336	D. Florescu (63)	0.1206	J. D. Ullman (82)
111 (163)	E. Rundensteiner	124 (222)	E. Bertino	0.1321	S. Sudarshan (68)	0.1173	P. A. Bernstein (72)
109 (192)	D. Agrawal	120 (178)	C. Faloutsos	0.1279	J. Hellerstein (73)	0.1159	M. Lenzerini (64)
109 (153)	M.J. Carey	119 (180)	G. Wiederhold	0.1271	H. Pirahesh (84)	0.1117	C.S. Jensen (79)
108 (147)	D.J. DeWitt	119 (148)	U. Dayal	0.1240	J. Widom (73)	0.1112	J.F. Naughton (89)

5. THE GIANT COMPONENT

The *giant component* of a graph is the largest subset of interconnected nodes in the graph. The rest of the nodes usually form much smaller components, typically of size *O(log n)*, where *n* is the total number of nodes [10]. In order to determine if such a component exists in our collaboration graph, we measured the relative size of the largest component which is simply the ratio of the nodes in the component to the all nodes in the graph. The growth of the giant component of our graph is shown in Figure 12 as a function of time.

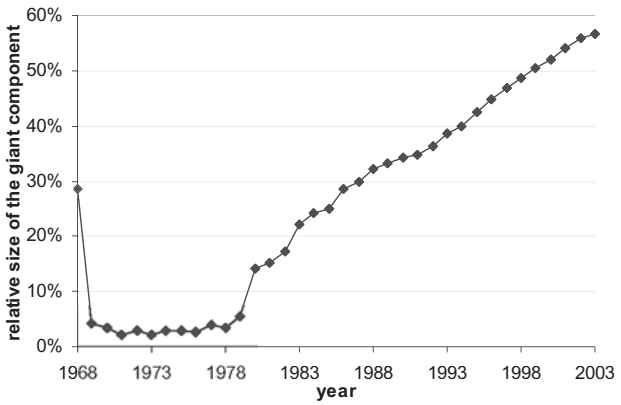

Figure 12: Relative size of the giant component, determined for the cumulative data up to the year indicated.

In the initial years, the size of the giant component of the graph is much smaller compared to the total number of nodes available in the graph, covering only about 5% of the whole graph although new authors keep joining to the db community. Yet, those authors help cluster other large components in the graph. In 1980, those clusters started to form larger components. After the size of the giant component exceeds 30%, it constantly increases up to the end of the period analyzed.

Although tendency for more collaboration in the community helps the smaller components to be connected to the giant component, the main increase stems from the new authors. In recent years the db community grows 10% by the addition of new authors, who tend to collaborate with existing authors (e.g., graduate students collaborating with their advisor) to write a paper instead of making a contribution alone. (Figures 1 and 7: the number of new authors, the number of single authors).

As of 2003, the size of the giant component is 18542, 57% of the whole db community. This is a rather low figure since the db community is expected to be a tight one. In addition, the second largest component is much smaller; it includes only 51 authors, who work on very particular subjects and publish mostly in 'Information Sciences' journal. The collaboration graph has 424 "isolated" components with 5-9 authors and 2892 components with 2-4 authors.

6. CLUSTERING COEFFICIENTS

Given a node *v*, the *neighborhood* of *v*, *N(v)*, is a subgraph that consists of the nodes adjacent to the node *v*. Furthermore, let us denote the edges and nodes in *N(v)* by *E(N(v))* and *K(N(v))*, respectively. Then, the clustering coefficient of *v*, $\gamma(v)$, is:

$$\gamma(v) = \frac{|E(N(v))|}{|E\max(N(v))|}$$

When the neighborhood is fully-connected (i.e., clique), it has $|E\max(N(v))| = \frac{|K(N(v))|(|K(N(v))|-1)}{2}$ edges. Therefore, the clustering coefficient measures how many edges actually occur compared to the fully-connected case [14]. The clustering coefficient of a graph *G*, $\gamma(G)$, is the average clustering coefficients of all nodes in *G*.

The clustering coefficient can be also viewed as "transitivity" which describes the interactions among trios of nodes in a network [10] – the degree to which a scholar's collaborators have collaborated with each other. In co-authorship networks, this measure implies how much authors are willing to collaborate with each other. The clustering coefficient of the giant component in the db co-

authors graph is shown in Figure 13 as a function of year (shown from 1972 when the network started to form a relatively giant component).

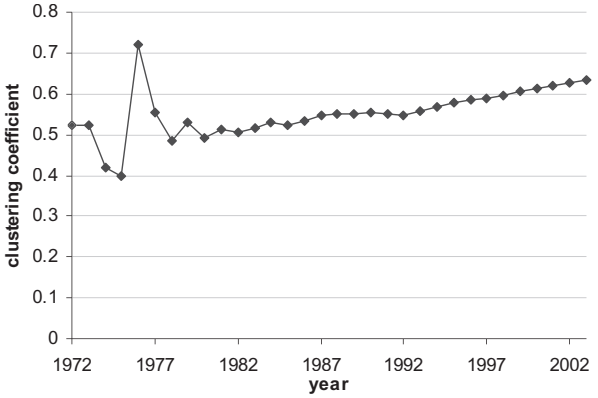

Figure 13: Clustering coefficient of the giant component, determined for the cumulative data up to the year indicated.

Over the years, the clustering coefficients tend to increase steadily, reaching 0.63 in 2003. This rather high value of the clustering coefficient is expected since DBLP-DB is after all a tight community of people working on databases only. Moreover, the increasing clustering coefficient is in sync with the tendency for more collaboration in recent years. The last two columns of Table 4 lists the top-10 authors with the highest clustering coefficients for different cases; the first ranking only considers authors with at least 60 collaborators while the second one lists authors with more than 60 papers (numbers in parenthesis).

7. GEODESIC

In a co-authorship network, two authors could know each other through their collaborators. In other words, there could be several interaction paths between two of them not directly but through a number of the other authors in the network. The path(s) with the minimum number of edges between any given pair of authors in the network is called shortest path or *geodesic* of the pair. Then the average distance in a network is the average of all pair-wise geodesics of authors in the network. Social networks often have small average distances compared to the number of nodes in the networks, which is first described by Milgram [7] and now referred to as "small-world effect". Figure 14 shows the evolution of the averages distance in the db community over the given period.

After the initial fluctuations, the average distance reaches to its maximum value, 8, in 1983. This seems natural since in the beginning of the time period analyzed, the growth of the community is rapid (i.e. each year, an increase between 40% - 80 % from the previous year). New authors are probably responsible for making the community expansion since the collaboration during that period was not very active. After 1983, however, it starts to decrease and eventually stabilizes around 6 for the last 15 years. Interestingly Milgram also found an average distance of six hops for his social network experiment. The relatively low value, 6, is probably a good sign since scientific discoveries can be disseminated rather fast [10].

The *diameter* of a graph, the maximum of the pair-wise distances in the giant component, of the db community is 20 as of 2003: "*A.Baczko - F.Seredynski - P.Bouvry - J.Blazewicz - P.Dell'Olmo - H.Kellerer - A.Caprara - D.Maio - P.Tiberio - S.J.Finkelstein - I.S.Mumick - O.Shmueli - F.Gavril - J.Urrutia - V.Estivill-Castro - L.Brankovic - M.Miller - P.D.Manuel - J.AlGhamdi - M.Sarfraz - K.Salah*".

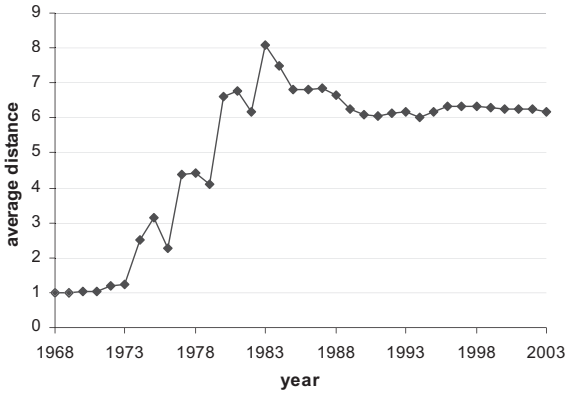

Figure 14: Average distance of the network, computed on the cumulative data up to the year indicated.

8. CENTRALITY

One of the interesting aspects of co-authorship network is to identify the most "central" scholars in the network. Authors who are the most prominent in the community are often (certainly not always) located in the strategic locations of the co-authorship network, which may allow them: (1) to communicate directly with many other authors, (2) to be close to many other authors, or (3) to be as intermediary in the interactions of many other pair of authors. There are several methods which aim to quantify authors' locations in [12]. For our work, we used the *closeness* and *betweenness* measures to quantify the prominent db scholars.

Table 5: Authors with the highest betweenness and closeness scores

Betweenness		Closeness	
0.054620	G. Wiederhold	0.268216	U. Dayal
0.048295	U. Dayal	0.262397	G. Wiederhold
0.045001	J. Han	0.256737	R. T. Snodgrass
0.038067	Y. Kambayashi	0.256555	D. Maier
0.030376	E. Bertino	0.256332	K. A. Ross
0.029406	H. Lu	0.256261	H. Garcia-Molina
0.027622	H.-J. Schek	0.256247	S. Ceri
0.026841	M. Jarke	0.256003	H.-J. Schek
0.026526	R. Agrawal	0.254401	M. J. Carey
0.026504	S. Ceri	0.253945	M. Stonebraker

8.1 Closeness Centrality

The *closeness* can be defined as how close an author is on average to all other authors. Authors with low closeness values could be viewed as those who can access new information quicker than others and similarly, information originating from those authors can be disseminated to others quicker [10]. Formally, the closeness of a node *v* in a connected graph G is defined as follows:

$$C(v) = \frac{n-1}{\sum_{v,w \in G} d(v,w)}$$

where *d(v,w)* is the pair-wise geodesic and *n* is the number of all nodes reachable from *v* in *G*. That means that, it is 1 over the average of the shortest paths from *v* to all other nodes in *G*.

The second column of Table 5 lists the top-10 individuals with the highest closeness scores. Furthermore, all those 10 scholars are closely connected to each other through collaborations (Figure 15), which could be viewed as a "core" component of the db community.

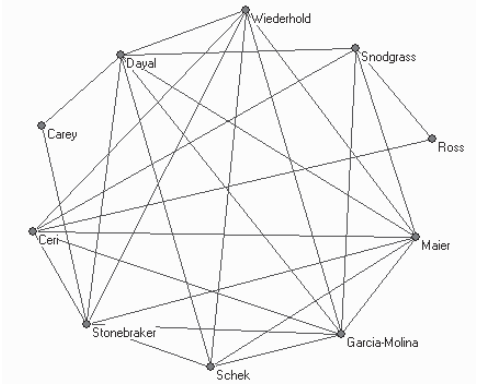

Figure 15: Connectivity of top10 authors with the highest closeness

8.2 Betweenness Centrality

Sometimes the interactions between any two non-directly connected authors (i.e., who never collaborated before) might depend on the authors who connect them through their shortest path(s). These authors potentially play an important role in the network by controlling the flow of interactions. Hence the authors who lie between most of the shortest paths of the pairs of authors could be viewed as the central people in the community. This notion, known as the *betweenness* of a node *v*, *B(v)*, measures the number of geodesics between pairs of nodes passing through *v*, and formally defined as follows [5]:

$$B(v) = \sum_{v,w,x \in G} \frac{d(w,x;v)}{d(w,x)}$$

where *d(w,x)* is a geodesic between *w* and *x*, and *d(w,x;v)* is a geodesic between *w* and *x* passing through *v*. The equation can be also interpreted as the sum of all probabilities a shortest path between each pair of nodes *w* and *x* passes through node *v*. The first column of Table 5 shows the top-10 authors with the highest betweenness scores.

8.3 Weighted Measures

So far, we have not differentiated whether or not two authors have single or multiple collaborations – as long as there is a single co-authored paper, two authors are linked in the collaboration graph. People have recognized this and tried to incorporate weight such that the more collaboration two authors have, the stronger link exists between them [1, 10]. Newman in [10] defines such a collaboration network as follows:

$$\sum_{j(\neq i)} w_{ij} = \sum_{k} \sum_{j(\neq i)} \frac{\partial_i^k \, \partial_j^k}{n_k - 1}$$

In this model, w_{ij} represent the collaboration weight between two authors *i* and *j*. ∂i^k is 1 if author *i* is a co-author of paper *k* and n_k is the total number of co-authors of paper *k*. That is, if authors *i* and *j* co-authored a paper *k*, each one should divide his time equally between the other *n-1* co-authors. Then, the sum of all collaboration weights w_{ij} between two authors defines the total strength of that tie.

Table 6: Closeness and betweenness for the weighted collaboration graph.

Weighted closeness			
score	**name**	**# of papers**	**# of co-authors**
0.19262862	H. V. Jagadish	106	102
0.19192969	Divesh Srivastava	85	91
0.19161295	Umeshwar Dayal	103	119
0.19156657	Raghu Ramakrishnan	90	76
0.19137470	Rakesh Agrawal	115	94
0.19029398	Surajit Chaudhuri	83	67
0.19005407	Jiawei Han	147	125
0.19004566	L.V. S. Lakshmanan	66	58
0.18976747	Hector Garcia-Molina	151	139
0.18934639	Qiming Chen	35	17
Weighted betweenness			
score	**name**	**# of papers**	**# of co-authors**
0.48422084	Umeshwar Dayal	103	119
0.46723451	Jiawei Han	147	125
0.33170557	H. V. Jagadish	106	102
0.28326387	Yahiko Kambayashi	105	99
0.28119593	Elisa Bertino	128	124
0.26476805	Hongjun Lu	97	99
0.26315179	Hector Garcia-Molina	151	139
0.25839473	Raghu Ramakrishnan	90	76
0.23657271	Surajit Chaudhuri	83	67
0.22207486	Rakesh Agrawal	115	94

We regenerated our network according to this model and considered the distance value between two authors as the inverse of their collaboration weight. The new weighted

rankings of closeness and betweenness measures can be seen in Table 6 for this network. Interestingly, authors who tend to collaborate often with a small number of people only are ranked high (e.g., scholars in research labs or a small set of collaborators).

8.4 Caveats

Although the bibliometric analysis described above is meaningful in many cases, it is worthwhile to point out that these are not without problems (Dickson even argues "measuring scientific productivity by tracking the publication record of researchers is widely acknowledged to be hazardous and imperfect" [4]). One such problem visibly occurs in our study as well.

Note that since the raw data set, DBLP-DB, contains only papers in the pre-selected database-related publication venues, all measures tend to favor authors whose expertise is focused on the database field alone (penalizing scholars whose expertise is diverse and inter-disciplinary). In addition, the measures such as closeness or betweenness cannot identify scholars who made a critical contribution to the community with only a small number of publications or collaborators.

For instance, consider the following four distinguished database scholars: (1) *E. F. Codd*: the inventor of the relational data model, (2) *Jim Gray*: Turing award winner, (3) *Peter P. Chen*: the inventor of the ER model, and (4) *Jeffrey D. Ullman*: a renowned computer scientist at Stanford University. As shown in Table 7, Chen and Codd are ranked very low (e.g., 347-th and 3638-th in their betweenness ranks) in the betweenness and closeness ranks due to their small number of publications. On the other hand, although both Ullman and Gray have a substantial number of publications (233 for Ullman and 121 for Gray), since their contributions in general are very diverse, ranging from algorithms and automata to databases and programming languages, and to even physics, only about 1/3 of them are included in DBLP-DB.

Table 7: Statistics of four authors for DBLP-DB. Figures in the parenthesis are for the entire DBLP data set.

	# of papers	# of co-authors	Betwn. rank	Close. rank
Ullman	82(233)	87(131)	67	17
Chen	33(43)	24(27)	347	837
Gray	43(121)	97(179)	83	29
Codd	23(47)	5(15)	3638	7829

9. CONCLUSIONS

In this paper we analyzed the collaboration network of scientists who publish in the database area. We presented a large number of statistics including how number of papers per author, authors per paper and number of collaborators change over the time period analyzed. We found that distributions of these statistics follow a scale-free power law distribution. We also looked at the evolution of other properties including average distance, clustering co-efficient and size of the giant component. The results imply that the db community seems to be a "small-world" by having a very small average distance between authors and being highly-clustered. These results may be helpful for further efforts on the db community such as modeling the network growth that may allow us to predict the approximate network behavior at any given time.

10. REFERENCES

[1] Barabasi, A. L. et al. *Evolution of the social network of scientific collaborations.* Physica A 311, 590-614 (2002).

[2] Batagelj, V. and Mrvar, A. *Pajek – A program for large network analysis.* Connections 21(2), 47-57 (1998). http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[3] DBLP*, Computer Science Bibliography.* http://www.informatik.uni-trier.de/~ley/

[4] Dickson, D. *The Virtue of Science by Numbers.* SciDev.Net (August 2003). http://www.scidev.net/Editorials/index.cfm?fuseaction=readEditorials&itemid=83&language=1

[5] Freeman, L. C. *A Set of Measures of Centrality Based on Betweenness.* Sociometry 40, 35-41 (1977).

[6] Lotka, A. J. *The frequency distribution of scientific production.* J. Walsh. Acad. Sci. 16, 317-323 (1926).

[7] Milgram, S. *The small world problem.* Psychology Today 2, 60-70 (1967).

[8] Nascimento, M. A., Sander, J. and Pound J. *Analysis of SIGMOD's co-authorship graph.* SIGMOD Record 32(3): 8-10 (2003).

[9] NetDraw*, Network Visualization Software.* http://www.analytictech.com/netdraw.htm

[10] Newman, M. E. J. *Who is the best connected scientist? A study of scientific coauthorship networks*. Phys. Rev. E64 016131(2001); Phys. Rev. E64 016132 (2001).

[11] Snodgrass, R. T. *Journal relevance.* SIGMOD Record 32(3): 11-15 (2003).

[12] Wasserman, S. and Faust, K. *Social Network Analysis.* Cambridge University Press, Cambridge (1994).

[13] Watts, D. J. and Strogatz S. H. *Collective dynamics of 'small-world' networks.* Nature 393, 440-442 (1998).

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