Bibliometric Landscape of the ACM Digital Library

Ergin Elmacioglu
Penn State Univ., USA
elmaciog@cse.psu.edu

Dongwon Lee
Penn State Univ., USA
dongwon@psu.edu

Mario A. Nascimento
Univ. of Alberta, Canada
mn@cs.ualberta.ca

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 [9] in 1967 who identified the so-called "*six degrees of separation*" phenomenon based on the results of 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, Erdös number). Inspired by some of the recent attempts to apply social network analysis to scientific communities [11, 3, 10], in this article, we analyze the collaboration network made of the researchers in the computing area at large, searching for interesting patterns underlying the computing community and their publication behavior.

Since ACM Guide [1] is a high-quality citation digital library that has a very good coverage on computing literature, we chose to use ACM Guide as the data set for our analysis of the computing community. In particular, we examined citation data in ACM Guide from 1950 to 2004, which contained about 609,000 authors and 770,000 publications.

We use a *collaboration network* (or graph) where nodes represent authors and edges between any two nodes exist if those nodes represent authors who have co-authored one or more papers (about 1.2 million edges). Note that ACM Guide itself does not have a notion of "unique key" such as DOI (Digital Object Identifier). Instead, it 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 try to minimize the effect this problem by conducting two experiments - one with full names ("John Doe") and the other with the first initial followed by the last name ("J. Doe") - and use these as the upper and lower bounds of the statistics. Since the results of the both cases appear to be parallel, we only comment on the results of that with full names. For the visualization of our network analysis, we used Pajek [4], a freely available social networks analysis tool for non-commercial use.

Basic Statistics

First, we present various statistical analysis related to the authors. Figure 1(a) shows the number of "new authors" (ones who publish a paper for the first time in the venues covered) as a function of year. There is a small community

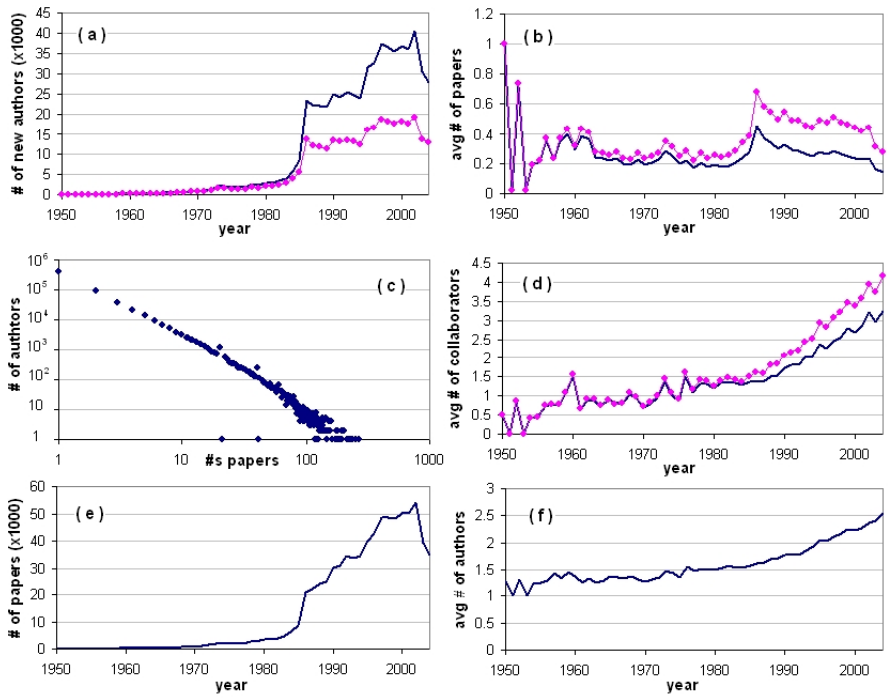

Figure 1: Basic Statistics (fig a-d: blue-solid lines for full name set, pink-dotted lines for abbreviated name set): (a) # of new authors per year, (b) avg. # of papers per author, (c) distribution of # of papers per author, (d) avg. # of collaborators per author, (e) # of papers per year, (f) avg.# of authors per paper.

in the first 30 years, this is due to the coverage of the database and less number of venues available then. In the late 1980s, the increase rate becomes around 40% and the community steadily grows. In the recent years, however, the same pace cannot be maintained, about 7% increase each year with slowing tendency. There are roughly 28,000 new authors in 2004 alone who publish a paper for the first time – novice graduate students or veteran scholars from related fields. Although the community is very large, each year, only a small portion (about 10%) of all authors are "active authors", i.e., ones who publish at least one paper in that year. Interestingly, almost half of the active authors in any given year are "new authors" and they are steadily contributing to about 55% of papers each year in the recent years. The remaining part is contributed purely by the existing authors, which increases the density of the network by creating new edges through new collaborations. Figure 1(b) illustrates the average number of papers per author for a given year. It does not seem to change dramatically, implying that the productivity rate of the computing community as a whole remains intact over time. This makes sense since only small fractions of the community are active each year and they can publish only a limited number of papers.

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*" [7]. He showed that such a distribution follows a power law with an exponent approximately -2. Figure 1(c) shows the distribution of numbers of papers per author on log-log scales for our database, of which the exponent is -2.59. Consistent with Lotka's Law, a small number of authors publish a large number of papers whereas 392,559 authors (64%) have only one paper (the fat tail on the right hand side indicates this). In fact, in ACM database, there are only 36 authors who published more than 150 papers. Top-10 authors with the most number of publications are shown in the first column of Table 1.

number of papers	number of co-authors	closeness	betweenness
266 B. Shneiderman	229 J. Dongarra	0.188847 S. Muthukrishnan	0.008837 B. Shneiderman
250 M. Karpinski	204 A. Gupta	0.187847 G. Wiederhold	0.008506 E. Bertino
236 H. Garcia-Molina	193 B. Shneiderman	0.187780 H. V. Jagadish	0.007474 L. T. Watson
234 P. S. Yu	181 I. Foster	0.187536 C. Faloutsos	0.007218 G. Wiederhold
226 M. Sharir	180 H. Garcia-Molina	0.187010 H. Garcia-Molina	0.007194 J. Dongarra
225 M. Stonebraker	175 E. Bertino	0.186284 E. Bertino	0.006921 M. Fujita
217 K. G. Shin	175 M. Stonebraker	0.185843 V. S. Subrahmanian	0.006799 S. Muthukrishnan
204 E. Bertino	174 G. Wiederhold	0.185719 M. Stonebraker	0.006697 W. Li
204 G. B. Shelly	174 R. Kikinis	0.185620 J. D. Ullman	0.006578 N. Alon
196 P. J. Denning	169 N. Alon	0.185590 U. Dayal	0.006246 C. Faloutsos

Table 1: The top-10 authors with the highest number of papers, number of co-authors, closeness and betweenness scores.

Now, we examine the number of collaborators an author has, illustrated in Figure 1(d). The average number of collaborators per author tends to increase steadily, 3.66 for the cumulative data up to 2004. Compared to other scientific communities having large-scale experimentation with large groups (e.g., high-energy physics), this average is rather small. 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, instant messaging, video conferences and the like) makes remote collaborations easier than before. The distribution of collaborators exhibits the power-law tail as well with exponent -2.84. The second column of Table 1 shows the authors with the largest number of collaborators.

The number of papers published each year (Figure 1(e)) shows very close relationship with the number of active (and new) authors verifying that the productivity is constant on average. In 2004 alone, there are more than 34,000 papers published, largely due to the increased number of authors. The average number of authors per paper tends to increase each year, yielding almost 2.6 co-authors per paper as of 2004 (Figure 1(f)). Although there are a significant number of papers with only a single author, the majority of papers are written by two authors (13557 papers). The figure clearly shows that there is an increasing tendency for collaboration among authors which also causes papers to have more co-authors.

Next, we looked at how publication venues are inter-related to each other using co-authorship information. By examining the pattern where the scholars publish their papers, one can see, for instance, which publications venues have a similar theme or orientation. Figure 2(a) shows a graph where (1) a node is a publication venue, whose size is proportional to the number of papers in it, and (2) an edge between venues X and Y reflects the similarity by the Jaccard distance, $\frac{|A \cap B|}{|A \cup B|}$, where A and B are author sets of venues X and Y. Hierarchical organization of the fields that have large venues with highly overlapping authors between each other is shown in Figure 2(b) while Table 2 lists the 10 pairs of the computing publication outlets having the highest similarity measure.

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 [11]. Figure 3(a) shows the relative size of the giant component in our collaboration graph, which is the ratio of the nodes

Similar Venue Pair		Distance
Journal of Algorithms	Symp. on Discrete Algorithms	0.19951
SIAM Journal on Computing	Annual ACM Symp. on Theory of Computing	0.18474
Computational Linguistics	Annual Meeting of the ACL	0.18042
Computational Geometry: Theory and Applications	Annual Symp. on Computational Geometry	0.18039
SIAM Journal on Computing	Symp. on Discrete Algorithms	0.17260
Annual ACM Symp. on Theory of Computing	Symp. on Discrete Algorithms	0.16362
Journal of Combinatorial Theory Series B	Journal of Graph Theory	0.16018
Journal of Computer and System Sciences	Annual ACM Symp. on Theory of Computing	0.15634
ACM TOPLAS	Annual Symp. on Principles of Prog. Lang.	0.15551
Journal of Functional Prog.	International Conference on Functional Prog.	0.15447

Table 2: Top 10 pairs of venues with the highest Jaccard distance.

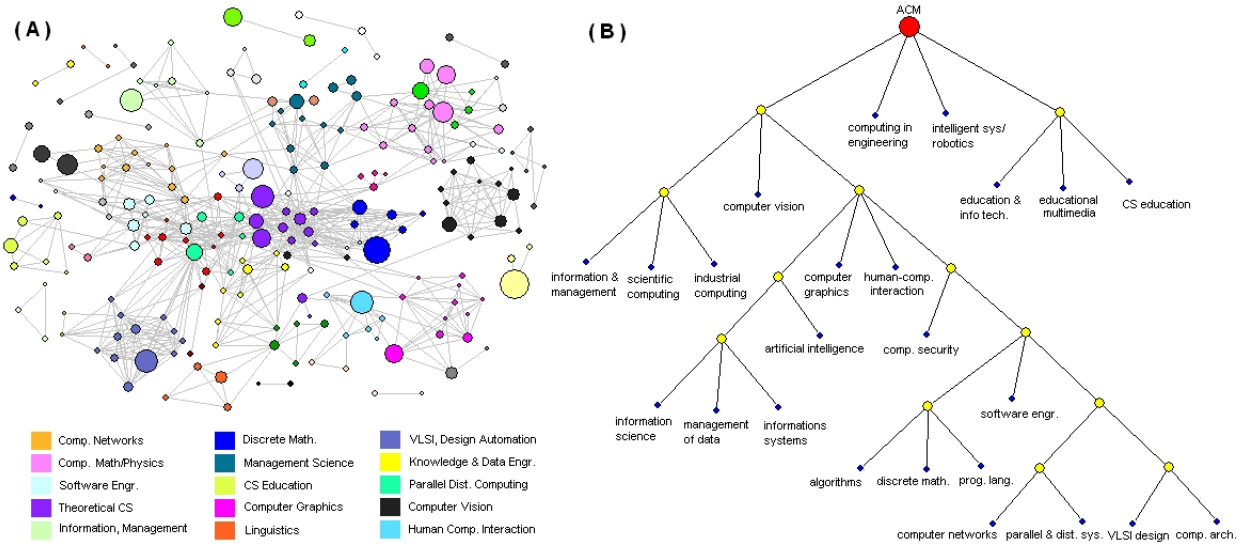

Figure 2: (a) Venue relations (only venues with at least 300 papers and edges with at least 0.3 Jaccard distance are shown). The size of a node is proportional to the number of papers in the venue. (b) Hierarchical organization of the fields that have strong relations with each other according to Jaccard distance.

in the component to the all nodes in the graph. In the initial years, the size of the giant component of the graph is much smaller compared to the total number of authors, covering only about 3% of the whole graph although new authors keep joining to the community. Yet, those authors help cluster other large components in the graph. In the mid 1970s, those clusters start to form larger components. The giant component then steadily grows till 2004. This is because new authors keep joining and existing authors collaborate more, smaller components are merged to the giant one.

As of 2004, the size of the giant component is 346137, 57% of the entire community. However, the second largest component is much smaller; it includes only 66 authors, who work on very particular subjects and publish mostly in "Cybernetics and Systems Analysis" journal. The collaboration graph also has 727 "*minor*" components with 10-19 authors and 5300 components with 5-9 authors. The reason for that many small components with remarkable number of authors might be due to: (1) geographical isolation of the research institutions since the database covers venues world wide, (2) variety of the subjects which have no overlap with the others, and (3) possibly name ambiguity problem.

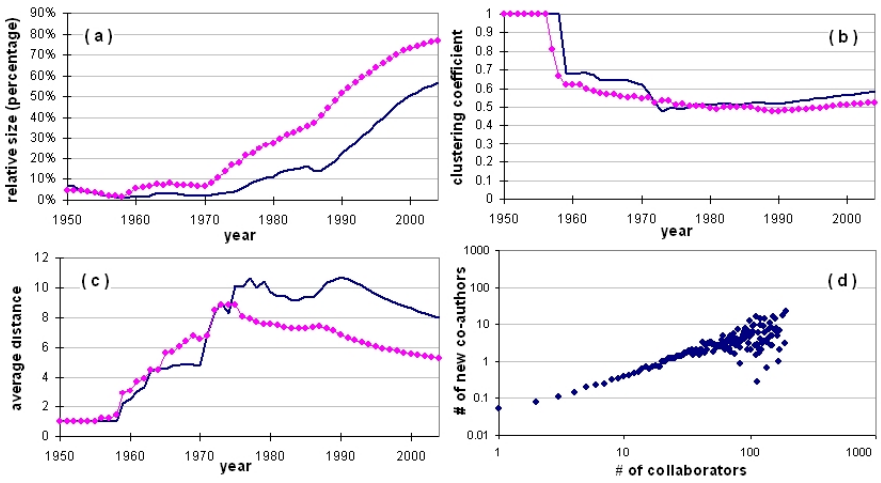

Figure 3: Structural Properties (blue-solid lines for full name set, pink-dotted lines for abbreviated name set): (a) relative size of the giant component, (b) clustering coefficients, (c) geodesics, (d) preferential attachment.

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:

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

where $|E_{max}(N(v))| = \frac{|K(N(v)|(|K(N(v))|-1)}{2}$ (when the neighborhood is fully-connected – clique) [12]. Therefore, the clustering coefficient measures how many edges actually occur compared to the fully-connected case. The clustering coefficient of a graph G, $\gamma(G)$, is the average clustering coefficients of all nodes in G. This measure could be viewed as the degree to which a scholar's collaborators have collaborated with each other.

The evolution of the clustering coefficient of the giant component is shown in Figure 3(b). First few years result in a fully connected graph; since the size of the component is small everyone has acquaintance with each other. Over the following years, the clustering coefficient tends to decrease as the giant component expands. Starting from the year 1974, when the giant component starts becoming relatively larger, the clustering coefficient shows a steady increase until 2004, mainly due to the increased collaboration among existing authors. As of 2004, it is about 0.6. This rather high value of the clustering coefficient is expected in such a social network.

Geodesics

In a co-authorship network, two authors know each other through their collaborators. The path(s) with the minimum number of edges between any given pair of authors in the network is called the shortest path or geodesic of the pair. Then the average distance in a network is the average of all pair-wise geodesics in the network. Social networks often have small average distances compared to the number of nodes in the networks, first described by Milgram [9] and now referred to as "small-world effect". Figure 3(c) shows the evolution of the average distances in the giant component of the community.

In the first forty years of the period analyzed, the geodesic tends to increase each year with occasional fluctuations. After it finally reaches its maximum, 10.7, in 1990, it decreases continuously because the increased tendency for collaboration creates new shorter paths between existing authors, prevailing the effect of expansion due to the new authors.

The final value of the geodesic in 2004 is 7.9 and seems to be decreasing in the following years ("8 degrees of separation"). Compared to the size of the community, this relatively low value is probably a good sign since scientific discoveries can be disseminated rather fast [11]. The diameter of a graph, the maximum of the pair-wise distances in the giant component, of the computing community is 33 as of 2004.

Centrality

An interesting bibliometric analysis of co-authorship network is to identify the most "central" scholars of the community. Authors who are the most prominent in the community are often (certainly not always) located in the strategic locations of the co-authorship network. *Closeness centrality* is one of the several methods which aim to quantify authors' locations. The *closeness* can be defined as how close an author is to all other authors on average. Authors with low values of this average 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 [11]. The third column of Table 1 lists the top-10 individuals according to the closeness scores, where the score of an author is normalized and defined as "one over the average value of the geodesics from that author to all others".

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 centrality* 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 fourth column of Table 1 shows the top-10 authors with the highest betweenness scores. The top authors in both rankings are indeed prominent scholars in the computing community.

Preferential Attachment

Preferential attachment, proposed by Barabasi and Albert [2], is an interesting phenomenon for modeling network growth. The basic assumption underling this phenomenon is that the probability that a node in a network gets new

Erdös #	Percentage	# of authors
1	0.01%	86
2	0.09%	571
3	0.72%	4376
4	4.33%	26348
5	13.54%	82502
> 5	38.13%	232291
infinite	43.18%	263028

Table 3: Distribution of authors' Erdös numbers

links is proportional to the number of links it already has, leading to a multiplicative process, which is known to give the power-law connectivity distribution. Due to its simplicity, several extensions have been widely used to model the growth of complex networks using only link structure [3], or both link and node properties [8].

For our social network, it implies that an author with many co-authors acquires new collaborators with a higher rate than one with less number of co-authors. In order to test if such a phenomenon exists, one needs temporal data, where the exact order that each particular collaboration is known. Then by making a histogram of the nodes with k links to which each new link is added, we have a function of preferential attachment. This should be an increasing function of k, believed to be a linear function in its basic form so that the resulting connectivity distribution (distribution of number of collaborators) will exhibit power-law [2]. Figure 3(d) shows the preferential attachment in our network. Since we have only data dated to the year, we present the change Δk in the number collaborators within one year interval (2000 - 2001), for an old author who had k collaborators at the beginning of the previous year (2000). Although the precision is not perfect, it exhibits preferential attachment behavior. The "rich" (ones with many co-authors) get "richer" (acquire more new co-authors than the others). In the case of no preferential attachment, the function would be constant, i.e., each author acquires new collaborators with the same probability.

Erdös Number

A well-known collaboration graph of mathematicians is available in "*Erdös Project*" [6]. In this graph, one distinguished node p represents a prominent mathematician "*Paul Erdös*", and the distance from each node v to p is known as of v's "*Erdös Number*". Thus, for example, Paul Erdös's co-authors have Erdös number 1, other co-authors of those have 2 etc.

As a final statistic, we present the percentage of the authors' Erdös numbers in our computing community using the data obtained from *Erdös Project*. Note that we only use the authors with Erdös number 1 and by locating them in our collaboration graph, we calculate the other authors numbers. In reality, Erdös number of an author might be different in case of a collaboration with some other author outside the computing community who has a smaller Erdös number. The results are shown in Table 3.

Conclusion

In this paper we analyzed the collaboration network of scientists who publish in the computing literature. 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 power law distribution. We also looked at the evolution of several structural properties. The results imply that the computing community seems to be a "small-world" having a small average distance with a value of 7.9 and being highly-clustered with clustering coefficient 0.6 for the largest connected component of the network which has 57% of all authors. These results may be helpful for further efforts on the computing community such as an accurate modeling the network growth that may allow us to predict the approximate network behavior at any given time.

References

[1] ACM Guide. http://portal.acm.org/guide/.

[2] Barabasi, A. L., and Albert, R. "Emergence of Scaling in Random Networks". *Science 286*, 509–512 (1999).

[3] Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and Vicsek, T. "Evolution of the social network of scientific collaborations". *Physica A 311* (2002), 590–614.

[4] Batagelj, V., and Mrvar, A. "Pajek - A program for large network analysis". *Connections 21(2)* (1998), 45–57. http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

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

[6] Grossman, J. "The Erdös Number Project". http://www.oakland.edu/enp/.

[7] Lotka, A. J. "The frequency distribution of scientific production". *J. Walsh Acad. Sci. 16* (1926), 317–323. http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

[8] Menczer, F. "Growing and navigating the small world Web by local content". *Proc. Natl. Acad. Sci. 99*, 14014–14019 (2002).

[9] Milgram, S. "The Small World Problem". *Psychology Today 2* (1967), 60–67.

[10] Nascimento, M. A., Sander, J., and Pound, J. "Analysis of SIGMOD's Co-Authorship Graph". *SIGMOD Record 32*, 3 (Sep 2003).

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

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