

“Publish or Perish” in the Internet Age

A study of publication statistics in computer networking research

Dah Ming Chiu and Tom Z. J. Fu
Department of Information Engineering, CUHK
{dmchiu, zjfu6}@ie.cuhk.edu.hk

This article is an editorial note submitted to CCR. It has NOT been peer reviewed. The author takes full responsibility for this article’s technical content. Comments can be posted through CCR Online.

ABSTRACT

This study takes papers from a selected set of computer networking conferences and journals spanning the past twenty years (1989-2008) to produce various statistics to show how our community publishes papers, and how this process is changing over the years. We observe the rapid growth in the rate of publications, venues, citations, authors, and number of co-authors. We explain how these quantities are related, in particular explore how they are related over time and the reasons behind the changes. The widely accepted model to explain the power law distribution of paper citations is *preferential attachment*. We propose an extension and refinement that suggests *elapsed time* is also a factor to determine which papers get cited. We try to compare the selected venues based on citation count, and discuss how we might think about these comparisons, in terms of the roles played by different venues, and the ability to predict impact by venues, and citation counts. The treatment of these issues is general and can be applied to study publication patterns in other research communities. The larger goal of this study is to generate discussion about our publication system, and work towards a vision to transform our publication system for better scalability and effectiveness.

Categories and Subject Descriptors

C.2.0 [Computer-Communication Networks]: General

General Terms

Documentation, Performance

Keywords

citation, h-index, g-index, co-authorship, academic publishing

1. INTRODUCTION

The academic publication machineries, taken as a whole, provides an archive for peer-reviewed academic papers. In the process, the meta-information associated with the publications, such as *date* and *venue* of publication, *authorship* and *citations* can also be readily gathered from various sources [1–5]. Such publication records can be very useful for research, though it is perhaps most often used in performance evaluation for hiring, promotion and tenure cases

in the academic world. In this paper, we study the publication records of a selected set of conferences and journals in the networking field in the past 10-20 years at an aggregate level, summarize, and discuss the statistics. Through these statistics, we hope to (1) share some interesting observations about the way our community publish, and the characteristics of some familiar conferences and journals; (2) ask questions and generate interest for future studies; (3) get feedback on methodologies and possible collaboration for this kind of studies.

For the rest of the paper, we begin by describing (a) the paperset we use in the context of existing on-line sources, and (b) the most important previous works and our approach. Subsequently, we discuss various results and observations one by one, ending with a conclusion section. For more detailed statistics and a detailed explanation of the methodology, see our technical report [9].

2. DATA MODEL

We begin with a brief definition of terminology and an explanation of the data we analyze.

A *paper* is published by a *venue* at a *date*. Venues refer to conferences (workshops) or journals. Most of them publish papers on a periodic basis (e.g. every year, or every month). A *paperset* is a collection of papers we use to calculate various statistics. Each paper has one or more *authors*; each author may publish one or more papers. As a rule, each paper cites related papers published earlier. The *citation count* of a paper, which changes as time goes on, is the total citation received by a paper by a given time. This data model is used by various providers of academic publication statistics, e.g. Google Scholar [1], DBLP [2], IEEE [3], ACM [4], ISI [5], CiteSeer [6], or Microsoft Libra [7]. These providers, however, tend to use different papersets. For example, CiteSeer, DBLP and Libra focus mostly on computer science and related literature, but each has its own rules of which conferences/papers to include or not. The various statistics, e.g. citation count of papers, are derived from the respective closed papersets, leaving them not cross-comparable. They also tend to use different metrics, e.g. ISI and CiteSeer may have different definitions of *impact factor* for venues.

For our study, we decided to focus on a particular research field, in this case, computer networking. We also decided to use the Google Scholar citation count (for each paper we consider) because Google Scholar’s paperset is arguably the largest. This decision has a couple of consequences: (a) The

process of gathering citation counts from Google Scholar cannot be completely automated. For a given set of papers, it takes some time to gather the citation counts with some level of manual verification. (b) The citation count of a paper in Google Scholar is continuously updated and changing. This means we need to focus on a limited set of conferences. In this study, we tried to choose those venues that we are more familiar with, and tried to pick different types to make them reasonably representative.

The paperset we consider comes from the set of venues listed in Table 1. We illustrate its relationship to the other papersets in Fig 1. The citation counts are gathered from Google Scholar in late summer of 2009, and can be considered as a snapshot.

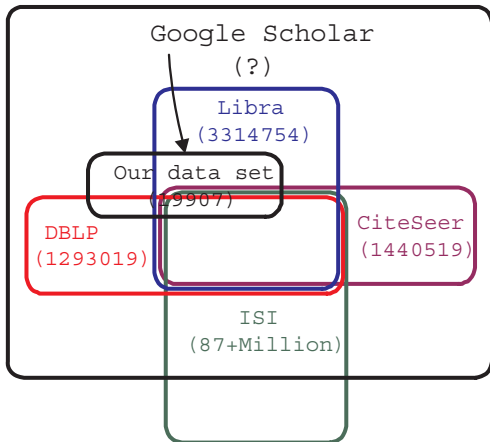


Figure 1: Illustration of several existing data sets.

Venue name	# years	Total # papers
Sigcomm (88-08)	21	612
Infocom (89-08)	20	4069
Sigmetrics (89-08)	20	795
Elsevier CN (89-08)	20	2953
IEEE/ACM ToN (93-08)	16	1331
ICNP (93-08)	16	559
MobiCom (95-08)	14	385
ICC (98-08)	11	7721
WWW (01-08)	8	1063
IMC (01-08)	8	281
NSDI (04-08)	5	138

Table 1: Paper set.

The study of academic publication statistics is by no means new. Previous attention focused mostly in different areas of science, especially physics. In fact, the field of this study is referred to as *Scientometrics*. The most influential work was published in 1965 by Derek de Solla Price, [10] in which he considered papers and citations as a network, and noticed the citation distribution (degree distribution) follows the power law. He tried to explain this phenomenon using a simple model, later referred to as the model of *preferential attachment* (i.e. a paper is more likely to cite another paper with more existing citations). Other authors also turned their attention to the network formed by authors who wrote papers together [13–17]. The difference in our study, other

than the focus on the computer networking field, is mainly in the consideration of how things change over time. In the discussion of our results, we will make passing reference to the prior works as the opportunities arise.

3. PAPER INFLATION

It should not be surprising to observe that we are seeing a rapid increase in the rate we are publishing papers in our field. We can account for the increase in terms of:

1. Many conferences increase the number of papers they publish, usually by increasing the number of *tracks* of presentation. Fig 2 shows the total rate of publication of the list of venues we consider.
2. Most successful conferences gradually add workshops before/after the main event, which allow more papers to be published. Fig 2 also shows the number of workshops associated with the list of conferences in our list.
3. The number of conferences and journals have increased significantly over the last 10-20 years. Fig 3 shows the number of venues in CiteSeerX’s *venue impact factor* report. CiteSeerX tracks a much larger field than networking; based on our experience, we are assuming the growth of the networking field is strongly correlated to that of the superset tracked by CiteSeerX.

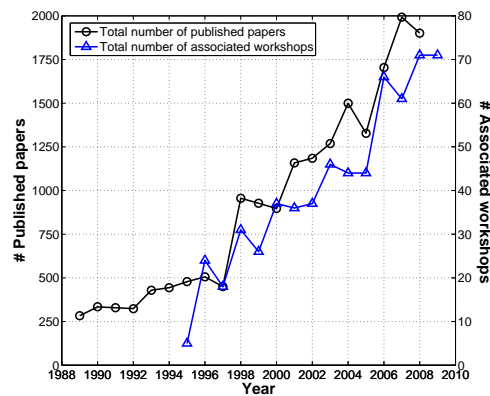


Figure 2: Total number papers and number of associated workshops changing by year.

Discussion - Likely reasons for paper inflation: We think the following are important reasons for paper inflation.

1. The number of authors is increasing. We will show some statistics in a later section on *authorship*. There may be various reasons for the increase, but two comes to mind: (a) The success of Internet and WWW has drastically raised people’s awareness and interest in working on computer networking. This is sometimes referred to as the *dotcom* effect. (b) Due the economic developments and opening-up of many developing countries, most notably China, a large number of authors are joining the research community overall.
2. The rate of publication per author is edging up. In order to raise the level of measurable output, many academic units or individual themselves strive to publish

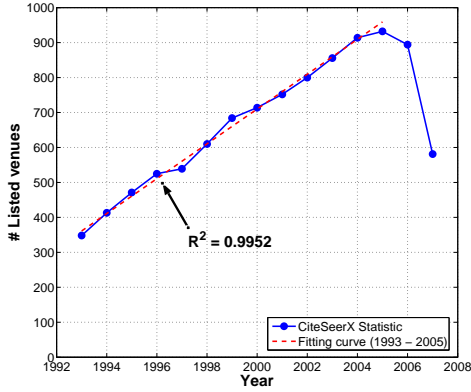


Figure 3: Number of listed venues by CiteSeerX in each year.

more. Sometimes a minimum number of publications is imposed before a student can obtain a graduate degree.

3. The Internet and WWW make publications more accessible. As a result, publications are shared more globally.

Discussion - Consequences of paper inflation: A rapid increase in publication puts great stress on the peer-review systems. Given the decentralized way publication venues are run, the relevant papers for any research topic may become more spread out, making it harder for researchers to follow the literature, hence increase the chances of *re-inventing the wheels*. There have been various proposals to revamp the whole publication system. It is indeed very exciting to think about how to design a global and scalable on-line system that all researchers can share their research results, and in the end all the accounting (of who did what) can be kept, and it remains highly usable (in terms of ease of finding the useful and relevant information quickly).

4. CITATION INFLATION

First, we plot the distribution of citation count earned by papers in our paperset using a log-log scale, as shown in Fig 4. The result is consistent¹ with prior works based on larger data sets [17].

It is more interesting to observe what happens when we plot the total citation count of time, shown in Fig 5. Preferential attachment alone cannot explain this, as there seems a pronounced dependence on time. The total citation count seems to be an increasing function till some point (7-8 years from the current time) then it starts decreasing.

This curve can be explained intuitively. First, the fall in total citation count in more recent years is due to the fact that citation count takes time to build up. The more

¹In order to ascertain the distribution indeed follows the power law, we can plot it as a Complementary Cumulative Distribution Function (CCDF). The result shows the data is closer to a lognormal distribution, which is also approximately heavy-tailed. In any case, the paperset in our study is a relatively small and biased data set. So the exact citation distribution is not the focus of this study.

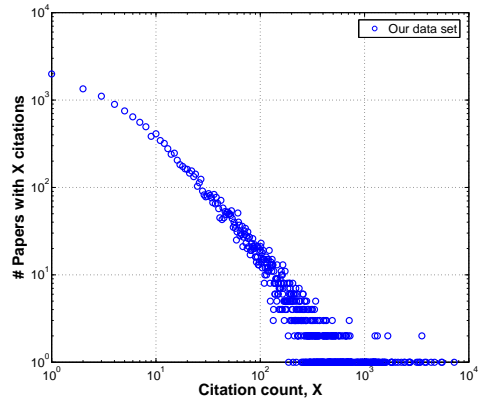


Figure 4: Distribution of number of citations (our data set)

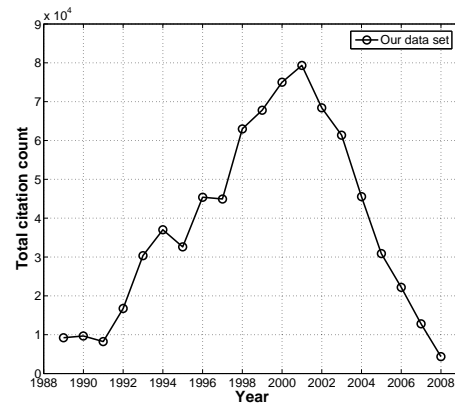


Figure 5: Total number of citations by year (our data set).

recent the year, the less the time for this build-up. The rise of the curve during the early years has also a good reason: it is a (subtle) consequence of paper inflation. By having fewer papers in earlier years, we lower the rate citations are collected in earlier years as well.

What is even more interesting is that we believe we discovered evidence that there is *fashion* in research. In other words, researchers may favor citing papers published in the recent past. This observation comes from our effort to create a model to explain the citation curve.

Let us try to explain the situation by a simple model. Let the number of papers published in year t be denoted by x_t , and the number of citations a paper makes be γ (a constant). The total number of citations made in year t is thus γx_t . The distribution of these citations to the papers published previously is assumed to follow a probability distribution $\alpha(n)$ where n is the number of years separating the citing paper and receiving paper. The sum of citations received by papers n years from the current year t is denoted $c(n, t)$. Assuming the time horizon for t is $[s, f]$, then

$$c(n, t) = \frac{\alpha(n)x_{t-n}}{\sum_{n=1}^{t-s} \alpha(n)x_{t-n}} x_t \gamma.$$

The total citations received in year t , denoted $c(t)$ is then

$$c(t) = \sum_{i=(t+1)}^f c(i-t, i).$$

We then use different plausible distributions for $\alpha(n)$, and try to fit the curve we have. We tried to use the uniform distribution and Poisson distribution with various parameters. The result for the Poisson distribution worked quite well. The paper increasing statistics is plotted in Fig 6. We used a smoothed version to model x_t . The dotted line is a projection into the next few years. The resulting citation curve predicted by the model (with a mean for α of 4 years²) is as shown in Fig 7. The *fashion* in research, based on this data set, is thus research topics first published 3-4 years earlier.

Finally, the dotted line curve is a prediction for inflation down the road - the expected number of citation counts in the coming years based on the trend.

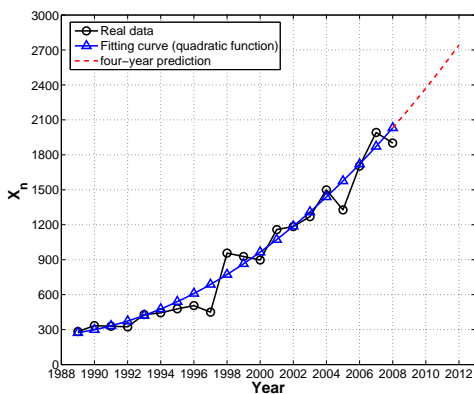


Figure 6: Total number of papers in each year versus quadratic fitting curve with four-year prediction.

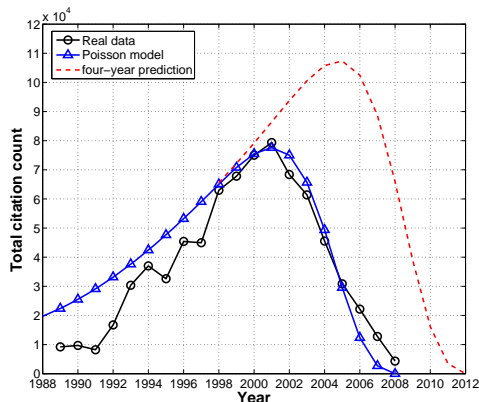


Figure 7: Total citation number of published papers versus Poisson model with four-year prediction.

Discussion - model assumption: Needless to say, our model assumes a closed paperset that cites papers internal

²Actually, the mean is three years, if we allowed citations for paper to be made to other papers in the same year.

to the set. In our analysis, our paperset is but a (small) subset of all the networking papers in the last 10-20 years. So there is inaccuracies due to papers in our paperset citing other papers, and other papers citing our papers. We are overlooking these issues in this simple analysis.

Discussion - adjusting to inflation: Citation count (and paper count) inflation is a phenomenon relevant to people making evaluations based on citation year after year. In year t , you may think a paper (or a person) receiving a citation count of at least c as adequate/good. In year $t + 1$, you would need to adjust your threshold up due to citation inflation. We have demonstrated how you might compute the inflation rate from a simple model, once you know about the paper inflation rate, and some idea of the citation distribution function α .

Similarly, when comparing two papers published at different times by citation count, you may consider using our model as a way to calibrate the comparison.

Discussion - research fashion? The model of *preferential attachment* alone does not seem rich enough to explain the aggregate behavior citation counts. The model of research fashion seems very interesting to us. To further validate the model, we need to use a paper database with more detailed information, e.g. including the times at which citations are made.

5. COMPARISON OF VENUES

Perhaps the most interesting question to readers is how the publication venues compare to each other, in terms of citation count. This corresponds to what is known as the *impact factor* in academic circles. There is no standard definition of impact factor, so we will use a number of different metrics to compute the citation statistics for different venues over a period of time: (i) The top twentieth paper; (ii) the average; (iii) the percentiles; (iv) the h-index; (v) the g-index. The results are tallied in Table 2.

H-index [8,12] and g-index [11] are recently proposed metrics for computing impact factor as a single number. To derive these numbers, you first sort all papers in descending order of citation count; h-index is the highest index (h) of a paper whose citation count is at least h , and g-index is the highest index (g) of a paper such that the sum of the citation counts from paper 1 to g is at least g^2 . H-index was originally proposed for evaluating the impact factor of a person. G-index is a variation of that. Both can be used to evaluate the impact factor of any aggregate of papers.

Comparing publication venues is controversial. There are several issues: (a) there is no well-established metric; (b) the venues publish papers at different rates (c) in our paperset, the data cover different time periods, and we know from the citation inflation discussion, it is not fair to compare citation statistics for paper published at different time. For (a), what we try to do is to apply many different metrics and let the readers make their own judgements. For (b) and (c), we re-computed all the statistics for a window of three years common to all venues. Indeed, the ranking for NSDI, the venue had the most negative bias, moved up in ranking in all metrics. The other conferences with 8 years of papers (WWW and IMC) also moved up slightly. The detailed results can be found in the technical report [9].

To remove the time-dependent effects, we can also try to do the comparison on a year by year fashion. Instead of doing this for all metrics, we do it using only one metric: plot

Venue name	# of paper	top 20th	Avg.	90-	80-	70-	60-	50-	40-	30-	20-	10-	h-index	g-index
Sigcomm	612	1155	233.5	513	281	181	114	79	58	34	16	6	179	362
MobiCom	385	962	201.3	484	230	142	93	62	39	22	12	5	124	276
ToN	1331	1026	99.4	206	102	61	39	24	16	10	5	2	167	333
NSDI	138	108	53.2	141	85	61	39	28	20	16	11	7	50	82
IMC	281	144	52.9	126	82	57	42	28	18	11	6	3	68	111
Infocom	4069	727	52.3	132	69	40	25	16	10	6	3	1	207	341
Sigmetrics	795	233	47.9	113	66	42	27	18	11	7	3	1	97	167
WWW	1063	293	40.8	120	51	32	20	11	7	3	2	0	110	176
ICNP	559	156	30.2	76	43	24	14	10	6	3	1	0	66	113
Elsevier CN	2953	453	29.2	54	25	15	9	6	4	2	1	0	127	241
ICC	7721	270	9.3	20	10	6	4	3	1	1	0	0	100	161

Table 2: Percentile- and index-based analysis on 11 selected venues.

the median citation count for *all* venues (in Fig 8). In this case, a couple of conferences consistently stand out - these are the ACM single-track conferences Sigcomm and MobiCom, with very low annual acceptance rates (about 10%).

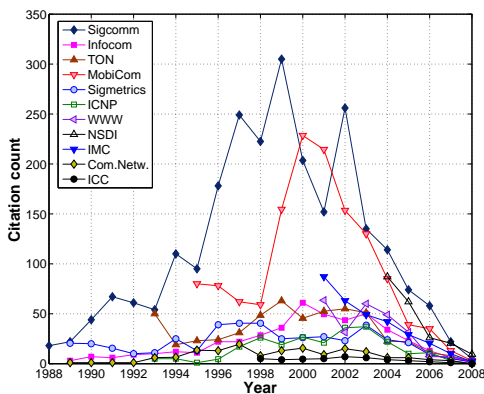


Figure 8: Median citation number of 11 selected venues in Computer Networking field with all paper counted in each year.

Discussion - different needs: First, it is expected that there are many levels of publications. In one extreme, we have papers based on new discovery, or a new idea, or by experienced researchers who have certain amount of self-discipline when submitting papers. In the other extreme, we have many papers written to mostly fulfill graduation, or job requirements. Different venues position themselves to satisfy different needs, in the process providing different kinds of service to the community.

Discussion - use of venue to judge impact: If the venue of a paper is used to predict the potential impact of that paper, the different venues have very different predictive power. For example, if we consider a (Google-scholar) citation count of 20 or higher to be of impact, then a Sigcomm paper is close to 80% likely to satisfy that threshold, whereas an ICC paper has only 10% likelihood of doing so. For venues with low predictive power, the citation count should be considered as well. In some academic institutions, only papers from journals are counted. This may be a reasonable simplification for conferences like ICC, but it is not wise to exclude those conferences that can indicate high

citation count reliably. For example, in our paperset, Infocom published comparable number of papers as the journal Computer Networks, and for each percentile, the paper from the former had more citations than that of the latter.

Discussion - use of citation count to judge impact:

Of course, citation count is not always a reliable indicator of impact. For citation count produced by Google Scholar, since it does not remove *self-citations* and it includes some on-line articles not peer-reviewed, it can have a lot of noise, especially when the value is lower than a certain threshold (of 10 for example). Finally, the correlation of the citation count of a paper to the quality and research value of a paper is complicated. Whether a paper gets cited seems to depend a lot on whether it gets introduced to and read by other researchers interested in the same topic, which can no longer be guaranteed these days since the rate of publication is higher than the ability of a researcher to follow them. Many other factors, such as presenting the paper to more people at different occasions, the prestige and the social connection and activeness of the authors, all seem to bias the citation count. To remove the effect of the noise, perhaps a condensed scale should be established (e.g. 0-10, 10-20, 20-50, 50-100, 100-200 etc counts as 0,1,2,3,4 and so on)³. The other major problem with the use of citation count is the initial time lag. For this, our model of deriving the effect of this lag on an average basis can be used as a rough predictor of citation count of a young paper. This is a topic worthy of further study. In our study, we do not have the citation history of individual papers. Some repositories provide this, such as ISI [5], but Google Scholar's output format does not make it easy to collect this information. It is interesting to study what the history (the rate of citation over time, and perhaps where the citations come from) can predict. For example, what type of impact? is it due to a controversy, or a short-lived hot topic? how broad and lasting is the impact?

6. CONFERENCE-THEN-JOURNAL

It seems we often publish a paper in conference, and then *journalize* it. How often do people try and successfully journalize a paper? It is said computer science people prefer to only look at conference papers, and usually do not bother journalize papers. We compiled some statistics to look at papers that were published in conference first and were journalized subsequently, versus papers that were not journalized. The method to determine a journalized paper is somewhat

³This method is used to score matches in contract bridge.

Conf.	Total	Journalized	Conf Cite	Jnl Cite	Total Cite	Non Jnl Cite
Sigcomm	541	108	276.5	298.9	575.4	193.1
Infocom	3415	597	34.9	82.1	117.0	56.0
Sigmetrics	691	88	37.3	84.3	121.6	47.7
ICNP	414	58	41.2	55.4	96.6	31.9
MobiCom	314	91	72.9	205.0	277.9	245.9
ICC	5547	336	7.5	17.8	25.3	12.3
WWW	598	65	71.9	97.6	169.5	58.1
IMC	209	17	82.6	74.2	145.8	60.4

Table 3: Comparison on average citation number between journalized papers and non-journalized ones.

ad hoc, so you need to take that into account when considering the results. For each paper in our paper set, we use Google Scholar to search for papers with approximately the same title (based on threshold of percentage match of regular expressions) and same authors (at least two common authors). The results are summarized in Table 3.

In this table, the second column is the total number of papers for each conference; the third column is the total number of journalized papers; the fourth column is the average citation count of those conference papers that are later journalized; the fifth column is the average citation count of those journal version of the conference papers; the sixth column is the sum of both conference and journal versions; and the last column is the average citation count of those papers that are not journalized.

Discussion - percentage of papers journalized: Across the board, less than 20% of the papers are journalized, and the overall average is significantly lower than 10%. The relatively low percentage may be due to different reasons. For ACM conferences, some authors may consider there is no need to journalize; so the limit may be caused by the authors. For other conferences, it can be limited by the quality of the paper.

Discussion - journalizing and citation count: First, again across the board, the total citation count (for the conference version plus the journal version), on average, is higher than those non-journalized papers (in last column). Apparently, two rounds of peer reviews can better sort out work with more impact - this is reasonable and expected. Note, this applies to the case of the ACM conferences as well⁴.

Second, the journal version, on average, receive more citations than the conference version. The exception is IMC. Our guess is that for a measurement conference, timeliness of a paper is more important than other type of papers, possibly leading to the poorer citation count for the journalized versions. This can also be due to a glitch, since the sample size for IMC is quite small. For this topic, we have done more analysis. For example, to find out the distribution of the number of years it takes to journalize (most often 1-2 years, but there are cases between 0 to 8 years); and to find out which journals the conference paper go. The readers are referred to read our technical report for more details.

Discussion - archival versus quality differentiation Originally, an important need for journalizing is because it serves an archival need. Nowadays, it can be argued that conference papers are equally well-archived. So the gain in journalization is mostly in allowing more time and more se-

⁴Although for Mobicom, the non-journalized papers scored quite high compared to the journalized cases.

rious reviewing to achieve higher quality. But in reality, the review process for journals are not that more extensive than the better quality conferences; and neither have consistent quality control. A scalable and consistent mechanism for quality differentiation of publications is a very interesting open problem.

7. AUTHORSHIP STATISTICS

We now turn to authorship statistics. Let us define a few quantities, each for a given period of time:

n : number of papers published;

m : number of distinct authors;

r : average number of papers published by each author;

q : average number of co-authors for each paper.

And there exists an invariant relationship connecting these quantities:

$$nq = mr$$

We can call this *the balancing equation of authorship*.

It can be easily verified that the equation is valid. There is one challenging problem - how to separate out distinct authors. We adopted the simplest solution - take the name from the paper *as is*. This method has two problems:

1. Name collision - more than one real person share the same name. They will be treated as the same author by us.
2. Multiple name representations - for example, Dah Ming Chiu is sometimes represented as D. M. Chiu. In this case, one real person is considered as two separate authors by us.

We will come back to this issue later in the section.

We can consider 1989-2008, the twenty years of our paper set, as one period. But the time aspect is lost. We are interested to see how the above statistics (n , m , r and q) changed over the twenty years. If we consider each year as a separate period of time, the sample size is small (hence variance would be large). So we divided the twenty years into four windows of five-year periods. The resulting quantities are tabulated in Table 4.

From the column for n , we see the number of papers published in successive 5-year windows increased very rapidly, as we discussed before. The number of distinct authors in the corresponding periods had equally rapid increases. Many of these authors (perhaps most students) wrote only a single paper, as shown in Fig 9.

Venue name	89-90	91-92	93-94	95-96	97-98	99-00	01-02	03-04	05-06	07-08
Sigcomm	1.90	2.27	2.45	2.47	2.78	3.01	3.12	3.51	3.63	4.35
Infocom	2.16	2.29	2.41	2.39	2.52	2.78	2.83	2.93	3.07	3.20
Sigmetrics	2.19	2.36	2.55	2.54	2.87	2.90	2.95	3.20	3.36	3.51
Elsevier CN	1.62	1.86	2.16	2.39	2.52	2.93	2.89	2.88	3.00	3.12
ToN	-	-	2.25	2.40	2.49	2.59	2.95	2.70	2.90	2.90
ICNP	-	-	2.53	2.82	2.63	2.62	3.00	3.05	3.26	3.32
MobiCom	-	-	-	2.66	2.87	3.07	3.21	3.34	3.36	3.41
ICC	-	-	-	-	2.57	2.69	2.66	2.80	2.92	3.05
WWW	-	-	-	-	-	-	3.17	3.50	3.07	3.35
IMC	-	-	-	-	-	-	3.19	3.25	3.81	3.57
NSDI	-	-	-	-	-	-	-	3.70	4.02	4.26

Table 5: Average number of co-authors per paper for each conference.

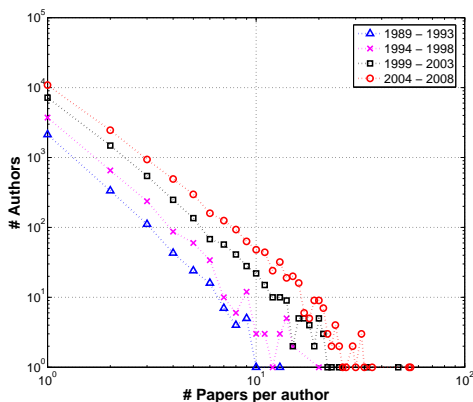


Figure 9: Distribution of number of papers per author (our data in 5-year window).

Window	n	m	q	r
1989 - 1993	1699	2678	2.161	1.371
1994 - 1998	2834	4843	2.489	1.457
1999 - 2003	5763	9873	2.817	1.644
2004 - 2008	9441	15731	3.074	1.845

Table 4: Window-based authorship analysis on our data.

It is interesting to note that the number of co-authors per paper has increased nearly 50%. At the same time, the number of papers per author also increased about 20-30%. These are all average numbers. If a large group of people maintain the same behavior (e.g. write only one paper), then the rest of the people must have incurred a much more significant change. In Table 5, we separately account for the number of co-authors per paper for each conference. We see that the number of co-authors stayed relatively more constant for some large conferences like Infocom. At the same time, some smaller conferences such as Sigcomm underwent much more significant increase in number of co-authors.

We found a way to double-check our method using another paperset. The DBLP [2] project open their paperset data for the public to use. Their paperset is much larger, containing approximately 1.1 million papers (1989-2008) covering a broader set of disciplines. We extract those published in computer networking venues in 1989-2008, a total of 105441

papers. For these papers, we performed the same (aggregate) authorship analysis. The results are shown in Table 6. If we plot the distribution of papers per author for each of the five-year intervals, the result is basically the same as that shown in Fig 9: We see increasing number of authors over time, but the distribution remains the same. This means a set of curves with the same (negative) slope, moved out slightly corresponding to the increased author numbers. In Fig 10, we plot the distribution for the entire period (1989-2008) for both papersets, for a comparison.

Window	n	m	q	r
1989 - 1993	13075	14657	2.066	1.843
1994 - 1998	17643	20446	2.234	1.928
1999 - 2003	24372	30252	2.458	1.980
2004 - 2008	50351	64781	2.840	2.207

Table 6: Window-based authorship analysis on DBLP data.

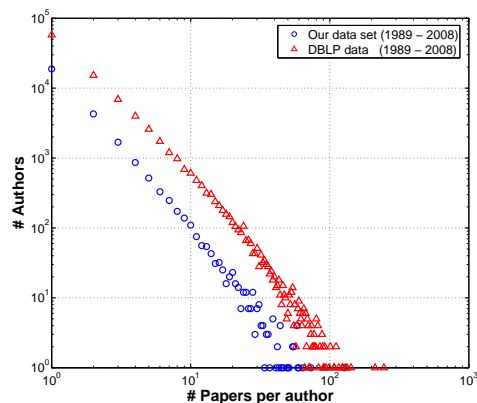


Figure 10: Distribution of number of papers per author (20 years of two data sets).

For the DBLP paperset, they made an effort to solve the *name collision* problem, by trying to distinguish authors based on their co-authorship patterns. The assumption is that two different real persons will have totally disjoint co-author groups. This does tend to eliminate name collisions, but it may also introduce another problem. When a person changes jobs, he/she is likely to build up totally disjoint

authorship relationships, and will be treated as two separate authors. Despite a different (and much larger) paperset, and a somewhat different way of identifying distinct authors, the resultant statistics are quite similar to those from our smaller paperset.

Discussion - authorship observations: These results on authorship confirmed our speculation on the reasons for paper inflation earlier. Indeed, we have all these causes - more authors, and authors are writing more papers on average. It behooves us to further study the type of authors and their contribution to the load on the publication system. Such information can be very helpful in the discussion of how to transform the current system into a more scalable system.

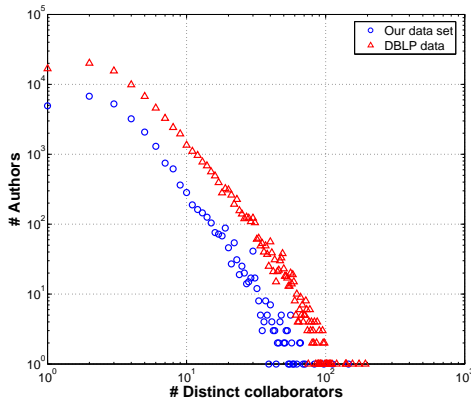


Figure 11: Distribution of number of distinct collaborators (two data sets).

Discussion - co-authorship implications: The co-authorship patterns and statistics can help us understand the prevailing collaboration trends, inter-institution or intra-institution. For intra-institution, the traditional mode may be dominated by student-supervisor collaboration. The increased co-authorship numbers may indicate group supervision, or hierarchical research groups are becoming more common. As pressure for more publications increase, there is also suspicion by some that some co-authors are free-riding. We are not able to ascertain this one way or the other.

Overall, the number of co-authors for computer networking (from our statistics) is similar to that for the field of physics (average 2.53) and biology (average 3.75), but higher than that for mathematics (average 1.45), as reported by a 2004 study by Newman [14].

Another interesting co-authorship analysis is to look at the collaborator distribution - how many collaborators an author has - on both data sets (in Fig 11).

As shown in Fig 11, the collaborator distribution results match against Newman's results (the physics and mathematics curves in [14]).

Discussion - relationship to other studies: Previous to this section, the analysis is based on only the paper citation network (node=paper, link=citation). For the authorship analysis, the network is extended to authors. There are two types of relationships: author-to-paper, and author-to-author (co-authorship). The author network is a special case of social network. There is considerable related work on this, and the more recent interest in on-line social net-

works. A proper survey of this area is beyond the scope of our current study.

8. AUTHOR PRODUCTIVITY

Suppose we agree to take the number of published papers as a measure of author productivity. We are interested to find out what factors have statistical correlation to productivity. We are able to study a few factors and report them here.

The first factor we consider is the *number of distinct collaborators*. The correlation (of number of papers to number of collaborators) is plotted in Fig 12. There is clearly some correlation.

Discussion - supervision of students: As we discussed earlier, there may be many patterns of collaboration. The collaboration between a supervisor and his/her students (or other hierarchical relationships) would lead a clear-cut correlation, a straight line with the slope corresponding to the average number of papers published involving a student. If other *lateral* collaboration relationships can be separated out, it would be interesting to see how the behaviour is different.

The next factor we consider is the *number of active years*. There are at least two ways to define active years: (a) the number of years, starting from the year of the first publication to the year of the last publication, of a given author; (b) the number of years in which a given author published at least one paper. We used the latter definition, and plot the result in Fig 13. Again, there is clearly positive correlation, as expected. But the diversity is quite high - from authors who publish one paper per year, to authors who publish about ten papers per year, on average.

Next we consider the correlation with the number of co-authors. The result is as shown in Fig 14. In this case, there does not seem to be noticeable correlation.

Finally, we show how a percentage of the most productive authors relate to the percentage of all papers published, in our paperset. In other words, we are interested in the minimum number of authors covering any percentage of papers published. This number can be computed approximately using a *greedy algorithm*, described as follows:

Initialization: Sort all the authors in a descending order of their productivity (number of published papers);

Repeat: Remove the author with most papers from the author list, and all his/her papers from the paper list; and plot (% authors removed, % papers removed).

The plot is shown in Fig 15, for both papersets (ours and the one from DBLP). Furthermore, we also plot (% authors, % citations), derived from the same greedy algorithm, replacing *papers* by *citations*. This is done only for our paperset, since citation information is readily available. It is interesting to note that both paper coverage curves satisfy the *20/80 rule*: twenty percent of the most active authors can take credit for eighty percent of the papers. The top 80% of the citations, however, can be credited to the top 5% of the authors.

Discussion - Partitioning authors: From Fig 15, it is perhaps reasonable to separate the authors into two broad classes: *professional* authors and *occasional* authors. For purpose of studying author productivity, we may see certain properties for the class of professional authors, which may

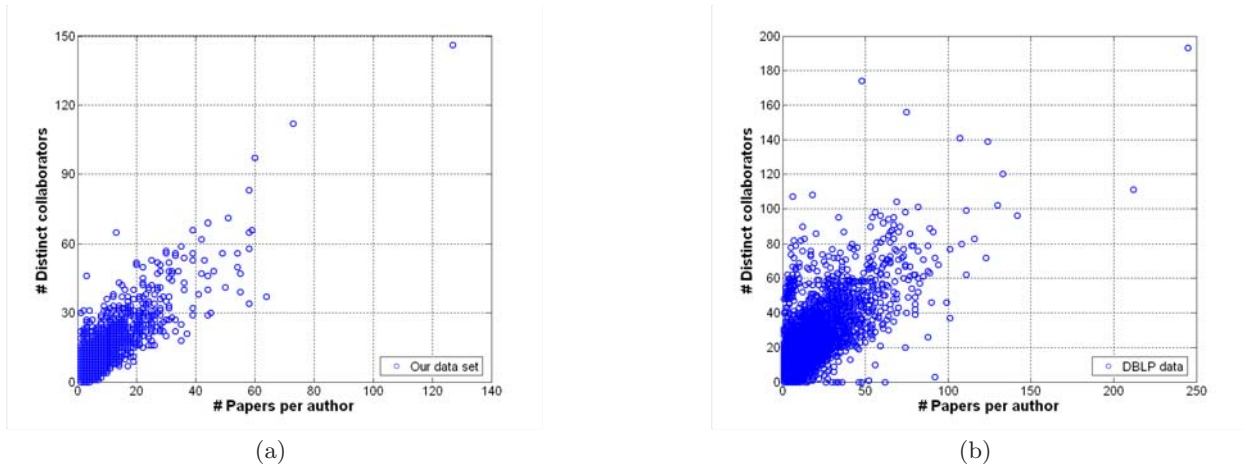


Figure 12: Productivity VS number of distinct collaborators (a) our data set, (b) DBLP data.

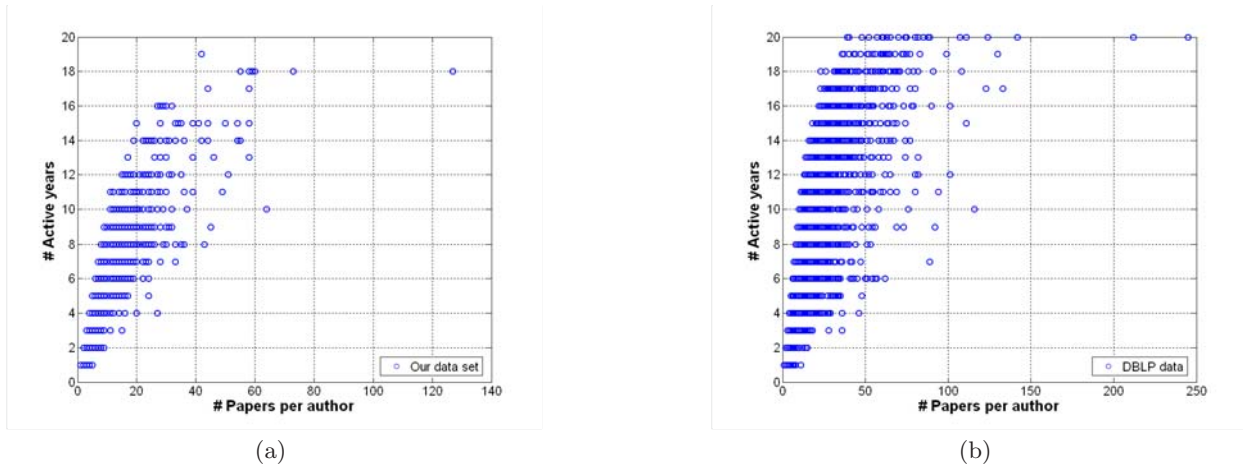


Figure 13: Productivity VS number of active years (a) our data set, (b) DBLP data.

not be easily visible when considering all authors together. This will be considered in future work.

9. RELATING CITATION COUNT TO CO-AUTHORSHIP

Finally, we study the relationship between the number of co-authors with citation count. In this case, we only used our own paperset. For the DBLP paperset, although they have the co-authorship information, the citation count information (based on Google Scholar) is not easily available. The result is plotted in Fig 16. The main figure is a scatter plot. In the inset, we also plot the average citation count for each number of co-authors. For higher number of co-authors, the sample size is very small, so the variance can be high. Between 1 to 7 co-authors, the average citation count seems almost flat, except for 5 authors. It is not clear if this is statistically significant, and if so why.

10. CONCLUDING REMARKS

What drove us to this work is a strong feeling that our publication system is running into a huge scalability prob-

lem. It seems we have endless number of deadlines for paper reviews and paper submissions, and yet we do not have enough time to read all the papers being published on the research problems we are working on. We also noticed how the number of publications and citations of researchers seem to be going through an inflation process, causing a lot of confusion on how to evaluate researchers for degrees and jobs.

We tried to pick a relatively small but representative set of conferences and journals and all the papers published at these venues in the past twenty years. We relied on available data and tools as much as possible, and tried to produce some meaningful statistics to our community, with discussion based on our perspective. We believe the model for citation history, and the balancing equation for authorship are useful tools and the time-based viewpoint is worthy of further studies. Our goal is to provoke more thinking by the whole community about how to improve our own system of publications, starting from how it is working now.

Acknowledgement

The authors wish to thank the student helpers, Kevin Lee and Ivy Ting, for their hard work and carefulness in data

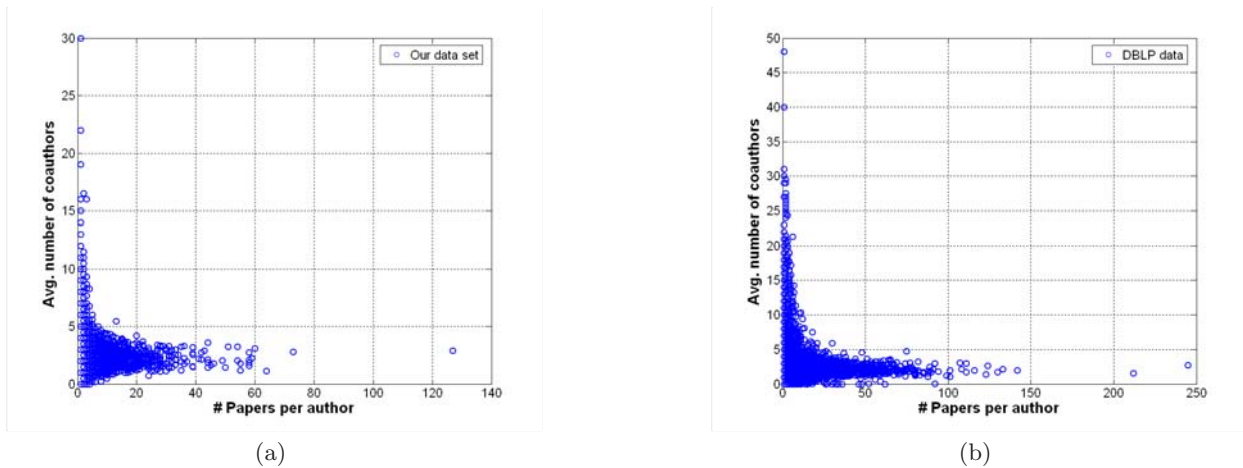


Figure 14: Productivity VS average number of co-authors (a) our data set, (b) DBLP data.

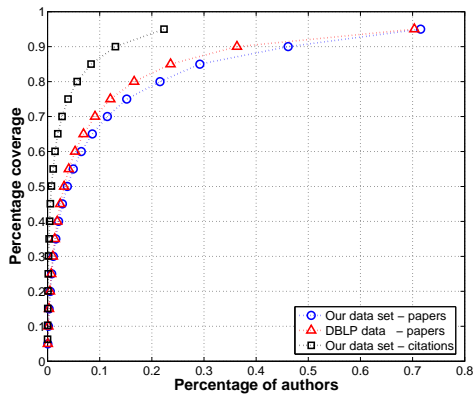


Figure 15: Approximate minimum number of authors for maximum paper and citation coverage.

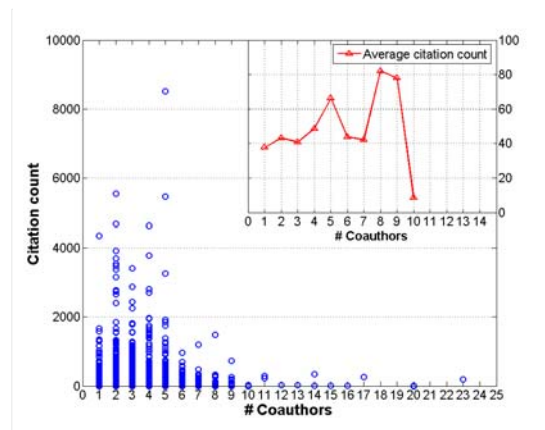


Figure 16: Correlation of citation count and number of coauthors.

collection; and thank Don Towsley, Vishal Misra and John Lui for various suggestions. The work is partially supported by a CUHK direct grant 2050411.

11. REFERENCES

- [1] “Google Scholar”, <http://scholar.google.com/>.
- [2] “DBLP”, <http://www.informatik.uni-trier.de/~ley/db/>.
- [3] “IEEE Digital Library”, <http://dl.comsoc.org/comsocdl/>.
- [4] “ACM Digital Library”, <http://portal.acm.org/dl.cfm/>.
- [5] “ISI Web of Knowledge”, <http://www.isiknowledge.com/>.
- [6] “CiteSeerX”, <http://citeseerx.ist.psu.edu/>.
- [7] “Microsoft Academic Search”, <http://libra.msra.cn/>.
- [8] P. Ball. Index aims for fair ranking of scientists. *Nature*, 436:900, 2005.
- [9] D. M. Chiu and T. Z. J. Fu. A study of publication statistics in computer networking research. Technical report, 2009.
- [10] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.
- [11] L. Egghe. An improvement of the h-index: The g-index. *ISSI Newsletter*, 2(1):8–9, 2006.
- [12] J. E. Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102:16569–16572, 2005.
- [13] R. K. Merton. The matthew effect in science. *Science*, 159:56–63, 1968.
- [14] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. USA*, 101:5200–5205, 2004.
- [15] M. E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Springer*, pages 337–370, 2004.
- [16] M. E. J. Newman. The first-mover advantage in scientific publication. *Europhys. Lett.*, 86:68001, 2009.
- [17] S. Redner. How popular is your paper? an empirical study of the citation distribution. *Eur. Phys. J. B*, 4:131–134, 1998.