INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Differences in Productivity and Impact across the Different Computer Science Subareas

J. Wainer *M. Eckmann* *S. Goldenstein*
A. Rocha

Technical Report - IC-12-08 - Relatório Técnico

March - 2012 - Março

The contents of this report are the sole responsibility of the authors.

O conteúdo do presente relatório é de única responsabilidade dos autores.

Differences in Productivity and Impact across the Different Computer Science Subareas*

Jacques Wainer[†] Michael Eckmann[‡] Siome Goldenstein[†] Anderson Rocha[†]

March 25, 2012

Abstract

Can we quantitatively compare different computer scientist? There is a widespread belief among computer science researchers that different subareas within Computer Science (CS) have different publishing practices (production throughput per year, preference either for journals or conferences, number of citations, etc.) making the use of a unified evaluation criterion unfair. In this paper we present productivity measures (both journal and conference productivity) and impact measures (citations per paper and H-index) for a random set of researchers in 17 different CS subareas. This research *quantifies* the mentioned intuitions and pre-empirical impressions and shows that, indeed, there are different publication practices and impact value measures for the different CS subareas. However, we show that few of the differences are significative.

1 Introduction

Some Computer Science researchers firmly believe that different subareas within Computer Science (CS) have different publishing practices, therefore a single production criterion would be unfair to some of the subareas. Indeed, as computer scientists ourselves, we have heard such comments on several occasions.

It is reasonable to believe that the subarea of Theory has different publishing practices than subareas such as Software Engineering or Image Processing. Scientific advances in Theory are often bounded by the time needed to prove theorems. Furthermore, at most institutions, CS

*The authors thank the financial support of CNPq, FAPESP, and Microsoft.

[†]The authors are with the Institute of Computing, University of Campinas (Unicamp), Av. Albert Einstein, 1251, Campinas, SP, Brazil, 13083-852. **E-mails**: {wainer,siome,anderson}@ic.unicamp.br

[‡]Dept. of Mathematics and Computer Science, 815 N. Broadway, Saratoga Springs, NY, USA. **E-mail**: meckmann@skidmore.edu

faculty working in Theory advise fewer students, and thus are likely to produce fewer publishable results per year.

It is also reasonable to believe that CS subareas that mainly deal with data, such as Image Processing, are more likely to have higher productivity than subareas in which the evaluation procedures require users (such as Human-Computer Interaction), programmers (such as Software Engineering), or organizations (such as Management Information Systems). The productivity of researchers in these human- or organization-based subareas are bounded by the difficulties of carrying out the empirical evaluations these fields require. These are all reasonable beliefs, however until so far that is all they are: unproven beliefs.

Besides the expected differences in productivity, we also often hear that different CS subareas (henceforth we will use the term CS *areas* instead of CS *subareas*) prefer and value differently conferences and journal publications. Bioinformatics seems to be more journal oriented, while Computer Architecture seems to be more conference oriented. We actually measure how much of each area's production goes to journals and conferences.

If indeed there are significant differences in publishing practices among the different CS areas, then a single production-based evaluation criterion for CS researchers will favor some areas and disfavor others. A probable consequence, beyond the unfairness to the disfavored areas, is that researchers will tend to avoid those areas in the future. Barbosa and Souza [1] discuss such a problem with respect to a uniform publication evaluation standard in Brazil and its negative impact on the Human-Computer Interaction area there.

Beyond publication practices, citation practices may also differ among areas. At least, areas with fewer researchers will probably reflect in lower citations to papers published in such areas. Again, uniform evaluation criteria for impact of one's research across the different CS areas will favor some areas and disfavor others.

The issue of evaluation of CS researchers has been discussed elsewhere [12, 11, 15], with special emphasis on the differences of scientific publication between CS and other scientific areas. Meyer et al. [11], for example, discusses the importance of conference publications in CS and states that, in some cases, a conference publication is more prestigious than a journal publication. The same general guideline of attributing importance to conferences is included in the Computing Research Association (CRA) guideline for faculty promotion in CS [12]. Wainer et al. [15] discuss that most of a CS researcher's work is not represented in the standard citation services such as Thomphson Reuters and Scopus (much less so than for Mathematics and Physics, for example), and thus when using metrics based on those services, a CS researcher or department may be unfairly evaluated, especially when in competition with other disciplines.

We are not aware of research that discusses the problems of uniform evaluation criteria across different CS subareas, except for [1]. In other scientific disciplines (e.g., Economics), there has been some discussion of the negative impact of uniform evaluation metrics for the different subareas of the discipline [8, 10].

General description of the paper

This paper's methodology relies on a sampling approach to evaluate the productivity and impact metrics of researchers in different CS areas, but we start with a definition of the different computer science areas. In Section 2.1, we discuss the areas chosen and our rationale for choosing them.

To define the researchers that work in an area we start with venues (conference and journals) that are clearly representative of each area — which we call the "seed venues" of each area (c.f., Section 2.2). From the seed venues, we use an iterative procedure that, given a set of venues representative of an area, computes the set of researchers that work in the area and then recomputes the venues. The process is repeated until convergence (c.f., Section 2.3). For the iterative process we use the DBLP data (obtained in July 2011) to define the universe of venues and researchers.

At the end of this step, we have a set of researchers that work in each area and a set of venues that are representative of each area. These are our populations. We sample 30 researchers from each area to evaluate the average productivity and H-index of the area, and we sample 100 papers published in the venues that represent the area to compute the average citations per paper.

In Section 2.4, we discuss the sampling of researchers and the productivity measure data collection. While in Section 2.5, we discuss the sampling of papers and the citation count data collection. Finally, we use the same sample of researchers per area to compute the average H-index as we discuss in Section 2.6.

2 Methods

2.1 Computer Science areas

When deciding how to define a CS area and which ones to select, we were guided by some of the classification schemes already in existence.

Both ACM and IEEE have area divisions (ACM Special Interest Groups (SIGs) and IEEE Technical Committees (TCs)) but some of them represent historical decisions that may not be as relevant in the present. The *DBLP*, *Scopus* and *Microsoft Academic Research* all have classified different CS areas, although none of them describes how they determined their specific classifications.

We want our set of areas to include both new and more traditional areas of computer science, to evaluate whether or not the traditional areas have different practices than the newer areas. Finally, we also want to include some areas on the fringe of CS that are not always present within the CS departments of institutions in different countries.

Table 1 lists the set of CS areas we chose. The table also lists the abbreviation we use for the area, which ACM SIG and IEEE TC group we believe corresponds to the area, and whether the area is included as a category (with a very similar name) in DBLP, Microsoft Academic Research, and Scopus.

The areas of Bioinformatics and Security play the role of the newer areas. Communications and Networking, Programming Languages, Databases, Computer Architecture, Distributed Computing, and Software Engineering are the more traditional areas. Operations Research and Management Information Systems are the two fringe areas — they have no corresponding IEEE TC, nor are they separated as an area in DBLP and Microsoft Academic Search.

The authors, who mostly work within the Computer Vision and Machine Learning areas, believe that Machine Learning is really a part of the general area of Artificial Intelligence but we acknowledge that there seems to be some strong evidence that practitioners in these areas see themselves as separate. ACM, IEEE, DBLP and Microsoft Academic Search all separate Machine Learning from Artificial Intelligence, so do we.

Of course, we **make no claim** that these are the only, nor that they are the most important areas of CS.

2.2 Seed venues

For each one of the areas, we select a small set of venues that “best” represents each area. The idea is that researchers working in each of the areas can clearly recognize such venues as “central” and “important” to their area. We call these the **seed** venues.

Seed venues should be central to each area: given that a researcher published a number of papers in these venues, one could, with high certainty, claim that the researcher works in that area. To get a sense of these “high precision” venues we asked colleagues that work in each of the selected areas. In some cases, according to these colleagues, the area itself was subdivided into smaller communities which may or may not have a large intersection. For example, the area of IPCV is composed of two sub-communities (Image Processing and Computer Vision) which may or may not have some intersection. For the cases that clearly have potential different sub-communities, we chose venues central to each sub-community.

Table 2 lists the seed venues for each area, in their usual abbreviation. Appendix B lists the full name of the venues.

Table 1: CS Areas — names, abbreviations, and corresponding ACM SIG, IEEE TC, and presence as an area in DBLP, Microsoft Academic, or Scopus.

Area	**Abbr.**	**ACM SIG**	**IEEE TC**	**DBLP**	**Microsoft**	**Scopus**
Artificial Intelligence	AI	SIGART	TCII	y	y	y
Bioinformatics	BIO	SIGBioinformat	TCCLS	y	y	
Communications and Networking	COM	SIGCOMM	TCCC		y	y
Compilers and Programming Languages	C+PL	SIGPLAN SIGAda	TCCL	y	y	
Computer Architecture	ARCH	SIGARCH SIGMICRO	TCCA TCDA TCuARCH TCVLSI		y	y
Computer Graphics	GRAPH	SIGGRAPH	VGTC		y	y
Database	DB	SIGMOD	TCDE TCMS	y	y	
Distributed Computing	DC		TCDP TCPP	y	y	
Human-Computer Interaction	HCI	SIGCHI			y	y
Image Processing and Computer Vision	IPCV		TCPAMI		y	y
Machine Learning	ML	SIGEVO SIGKDD	TCPAMI	y	y	y
Management Information Systems	MIS	SIGMIS				y
Multimedia	MM	SIGMM SIGWEB SIGDOC	TCMC	y	y	
Operational Research and Optimization	OR					y
Security	SEC	SIGSAC	TCSP	y	y	
Software Engineering	SE	SIGSOFT	TCSE		y	y
Theory	TH	SIGACT	TCCX TCMF	y	y	y

2.3 Iterative definition of venues and researchers in each area

In this section we describe the method we devised to find automatically which are the researchers and venues of each area. This allows for a methodical process that depends only on the seed venues and puts less emphasis and dependence on initially assigned venues and authors to areas. For this step we use data from DBLP (downloaded in August 2011) as the universe of researchers and venues of interest. DBLP[1] is a bibliographic server with focus on Computer Science which, according to its page, indexes over 1.8 million articles in CS.

At the start, Step 0, the set of venues that represent an area j is exactly the seed venues of

[1] http://www.informatik.uni-trier.de/~ley/db/index.html

Table 2: Seed venues for each CS area.

Area	**Venues**					
AI	AIJ	JAIR	JAR	AAAI	IJCAI	
ARCH	ISCA	MICRO	DAC	ASPLOS	TCAD	SC
BIO	BMC Bioinf	Bioinformatics	JCB	RECOMB	TCBB	
COMM	TON	TCOM	Mobicom	Sigcomm	Infocom	
C+PL	OOPSLA	POPL	PLDI	TOPLAS	CGO	
DB	TODS	VLDB	Sigmod			
DC	TPDS	JPDC	ICDCS	ICPP		
GRAPH	TOG	CGA	TVCG	SIGGRAPH		
HCI	TOCHI	IJMMS	UMUAI	CHI	CSCW	
IPCV	IJCV	TIP	CVPR	ICIP		
MIS	ISR	MANSCI	JMIS	EJIS	MISQ	
ML	JMLR	ML	NECO	NIPS	ICML	
MM	MMS	TMM	IEEEMM	MM	ICMCS	
OR	Math Prog	SIOPT	C&OR	Disc. Appl. Math		
SE	TSE	TOSEM	ICSE	TACAS	ESE	
SEC	TISSEC	JCS	IEEESP	SP	USS	CSS
TH	JACM	SICOMP	STOC	FOCS	SODA	

each area, denoted as V_j^0. At Step 0, the researchers that work in the area j, denoted as R_j^0, are the set of all researchers that, according to DBLP, co-authored at least two papers in any of the venues in V_j^0 from 2006 through 2010.

Note that a single researcher may work in two or more different areas. We call these researchers *multidisciplinary* researchers. In this case, it suffices that the researcher has co-authored at least two papers in two seed venues of each area.

At Step t, the set of venues that represent the area j is denoted by V_j^t, and the set of researchers that work in the area is denoted by R_j^t. At Step $t+1$ the set of venues for area j is defined as $V_j^{t+1} = V_j^t \cup \{v\}$ such that:

- $v \notin V_x^t$ for all x
- n_x^t is the number of researchers in R_x^t who co-authored any paper (from 2006 to 2010) in v
- then $n_j^t > n_k^t \geq n_{x1}^t \geq \dots n_{xn}^t$
- and $\frac{n_j^t - n_k^t}{\sum_i n_i^t} \geq \delta$.

We add venue v to the set of venues for area j if, among the authors of papers in v, researchers from j are the most common, and the ratio between the number of authors from j to the second most common area (k) is at least δ. The δ threshold is the minimum ratio of the relative diference between the difference of the number of authors in that venue from the area the venue was assigned to, and the number of authors from the second most common area. The value of *delta* was taken from the from the seed venue that was the most multidisciplinary. In this case, *delta* is 0.40, which is the relative difference between the number of DC researchers and COMM researchers that published in the conference ICDCS (a seed venue for DC).

At Step $t+1$, the set of researchers working in the area j are those researchers that:

- co-authored at least one paper p_1 in the set V_j^{t+1}
- co-authored at least one paper $p_2 \neq p_1$ in the set V_j^0

That is, we say a researcher works in an area if he/she co-authored at least one paper in any of the venues (at Step $t+1$) of the area, and at least one paper in the seed venues of the area. Finally, we only consider venues with at least 200 papers published from 2006 through 2010.

This methodology to simultaneously classify venues and authors is a contribution of this paper. We use a semi-supervised algorithm that used co-publication as the link between two venues. Most classification systems for venues are based on unsupervised algorithms, based on different citation links between the venues (for example [3]. Other research use also unsupervised algorithms based on co-authorship and topic detection on the documents themselves (for example [14]).

2.4 Productivity metrics

We exclude researchers who we have found to work in more than three areas because these often are homonyms (different researchers with the same name being grouped together). With the list of all researchers working in each area (including those who work in at most two other areas), we "randomly" select a subset of 30 of them and verify that each researcher in the sample has a webpage that lists all of his or her publications. Because this adds some "non-randomness" to the sample, we now describe the details of the sampling.

We randomly order the list of all researchers in the area, and go through this list in order, verifying if the researcher has a personal or institutional webpage containing his or her list of publications. For that, we search Google with the researcher name and current affiliation. If no personal or institutional page is found in the top 10 returned results, we consider that the researcher has no publication page, and move to the next researcher in the area. If we find a personal or institutional page, we search for a complete list of publications. If the page explicitly mentions "selected publications" we discard the researcher and move to the next one in the list. We repeat this process until we find 30 researchers for each of the areas with a complete list of publications.

This method adds some non-randomness to the sample because, we discovered, it is more likely that a non-senior faculty researcher in a western country university would have an up-to-date publications page, than the alternatives: eastern country researchers, students, industry based researchers, and senior faculty researchers.

For the 30 researchers per area, we collect the number of conference papers (including workshops but excluding posters) and journal papers the researcher lists on his or her page for the period from 2006 through 2010. We do not include in the analyses any non-English publications. We also collect the years for the first and last publications listed on the researcher's publication page and we determine the intersection of this interval with the interval from 2006 through 2010. This results in the **windowed publication interval** of the researcher. The researcher productivity is the number of journal and conference papers published between 2006 and 2010, divided by the windowed publication interval.

2.5 Citations per paper

For each area, we select a random sample of 100 papers from the authors published in the venues of the area, for the year 2006 and collected the number of citations received by these papers using Google Scholar. We collected this data in November, 2011. Given that citation counts are susceptible to outliers, we use the median citations per paper to perform the calculations.

2.6 H-index

To calculate the H-indices, we use the same sample of 30 researchers per area we discussed in Section 2.4 and collect the H-index of the first 20 of them for which it is feasible to collect the H-index data, as computed by the Google Scholar-based Publish and Perish tool[6]. Google Scholar will return all publications by researchers with the same last name as the researcher queried, thus if the researcher has a common last name, the number of publications returned is too high – we consider that collecting the H-index for these cases is unfeasible. In order to allow for some number of unfeasible cases, we drop the sample size from 30 to 20. Given that we have the list of publications of those researchers, we verify that only papers truly co-authored by the researchers are included in the H-index computation. For such analysis, we do not consider patents. The H-index calculation includes all papers published by the researcher, not only the papers in the 2006 through 2010 period.

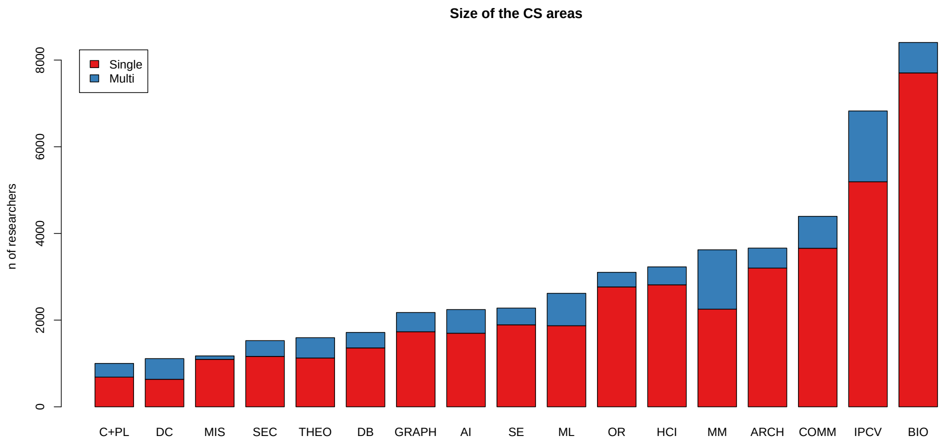

Figure 1: Number of researchers in each community. *Single* area denotes the researchers that only work in the area while *Multi* area denotes researchers that works in other areas as well.

2.7 Statistical Analysis

Given that we are comparing many sets of measures (one for each area) we must be conscious of the problems of multiple comparisons. Therefore, to evaluate the statistical significance of the comparisons, we use a pairwise t test (or a pairwise Wilcox rank sum test for the citation data) with the Holm correction [7]. We use a 95% confidence level to make claims of statistical significance.

3 Results

The iterative algorithm we discussed in Section 2.3 normally converges with one iteration. We have a set of 612 venues (247 journals and 365 conferences). Appendix B lists all the venues associated with each area.

3.1 Size of the research communities

We remove 226 researchers from the universe of researchers, because they work in more than three areas. There are 51,762 researchers that work in a single area, while 4,601 work in more than one area.

Figure 1 (and Table 5 in Appendix A) shows the size of each research community we consider in this paper. The data *Single* denotes the number of researchers that work only in that area while *Multi* denotes researchers that works in other areas as well.

3.2 Characteristics of the samples

Table 3 shows some of the characteristics of the sample of 30 researchers per area. The column *Students* reports the number of students (in 2010 or in the last year of their publication window) in the sample of 30 researchers. The column *Faculty in Non-CS Depts.* reports how many researchers in the sample are faculty in non-CS departments. We consider a non-CS department any that does not contain the words "computer" or "information" in their names. Thus a researcher whose webpage states that he or she is affiliated with an "Electrical Engineering" department (or college, or faculty) is considered as a non-CS faculty, but a researcher in an "Electrical and Computer Engineering" department was considered a CS faculty.

The areas of BIO, IPCV, MIS, and OR are the most "non-central" among the areas, based on the number of non-CS faculty in the sample. We were expecting that for the BIO, MIS, and OR areas, which were included in our set of areas exactly because they sometimes but not always are present in CS departments. For BIO, some of the non-CS faculty are hosted in Biology-related departments, for MIS, in Business and Marketing departments, and for OR in Applied Math and Engineering departments. However, surprisingly this is also true for IPCV researchers with about one third of them working in non-CS departments. In fact, we verified that most of the non-CS faculty in IPCV are hosted in Radiology and Medical departments.

The number of students found in the sample may indicate, in general, how attractive each of the areas are to new students. ARCH, DB, HCI, and MM seem to be the most attractive, while MIS is the least. (No statistical analysis was performed due to the low numbers).

3.3 Productivity measures

Figure 2 (and Table 6 in Appendix A) depicts the average conference and journal productivity of the sampled researchers in each area, in papers per year, ordered by total production. Appendix C.1 to C.4 present the statistical significance results.

Regarding total productivity, although the data seems to show three groups: the higher productivity group (ARCH, COMM, DC, and IPCV), the middle group, and a lower productivity group (MIS and OR), almost all differences *are not* significant at 95% confidence level). The only significant differences are, in general, between MIS and OR and the higher productivity group. There are also significant differences between either DB or TH and some of the higher productivity areas, but not all. Given the p-values of the pairwise comparisons, it is unlikely that larger samples would reveal that the differences between the middle group and the higher productivity group are significant. Thus one cannot claim that in general there are total productivity differences among the CS areas, except for few cases.

Table 3: Characteristics of the samples in each area.

Area	**Students**	**Faculty in Non-CS Depts.**
AI	6	0
ARCH	10	3
BIO	4	11
C+PL	4	0
COMM	4	0
DB	7	0
DC	3	0
GRAPH	4	3
HCI	8	3
IPCV	4	12
MIS	0	19
ML	4	5
MM	9	0
OR	2	19
SE	4	2
SEC	4	0
TH	3	4

For journals, BIO has a significantly higher productivity (3.44) than all other areas except the next two higher, COMM and ML. And COMM and ML are significantly different than the lower journal productivity areas of AI, C+PL and DB. No other differences are significant. The highest two conference productivity areas (DC and IPCV) are significantly different than the botton third (MIS, OR, BIO, TH, ML, and DB), whereas the lowest two (MIS and OR) are significantly different than the top third (DC, IPCV, ARCH, COMM, HCI, and SE).

Beyond productivity, if we consider the ratio of journal production to total production, there are basically two groups. On one hand, BIO, MIS, and OR, prefer journal publications over conferences with about 70% of their production oriented to journals. The differences to all other areas are significant. ML and TH represent an intermediary group, which publish almost half of its production in journals, and their differences to most of the other areas is also significant.

3.4 Impact measures

Figure 3 (and Table 7 in Appendix A) depicts the mean and median citations per year for the sample of 100 papers (from 2006) from each area (ordered by the median citations per year.) Appendix C.5 presents the statistical significance results.

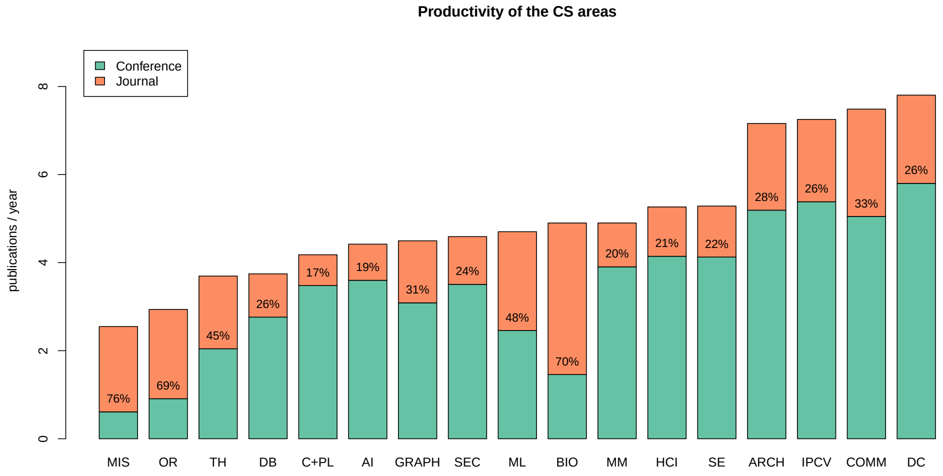

Figure 2: Productivity per year, ordered by total productivity. The numbers in each bar is the proportion of journal publications to the total production.

MIS citations per paper per year are not significantly different than the next four higher rates, GRAPH, DB, BIO and HCI, in decreasing order. The two lower rate areas, ARCH and MM are significantly different than the third lower rate area, DC. The other areas are in the same group with no significant differences among them.

The citation figures show an interesting relation to productivity. The higher productivity areas are the ones with lower citations rates. Thus, on the surface, in areas such as ARCH and MM, researchers are writing many papers each year, but few people are reading them (or better, citing them). Since MIS, the area with the highest citation rate is also the area with more emphasis on journal publication, it may be the case that an area's inclination to publish in conferences, given the usual restrictions on the number of pages of the publications, forces the authors to only cite a few of the relevant papers. There is indeed a positive correlation between citation rate and proportion of publications in journal, as shown in Figure 4, but the positive correlation is not enough to offset the negative correlation between citation rates and total productivity (Figure 4). A linear regression of citation rates and both total productivity and proportion of journal publication reveals that only the negative coefficient of the total production is significant (Appendix C.6).

The larger differences between mean and median citation rates for BIO, COMM, C+PL, and SE seems to indicate that in these areas there is an even higher than usual concentration of citations in a few papers (which would increase the mean but not the median). In fact the papers with the two largest citation counts in our sample are from BIO.

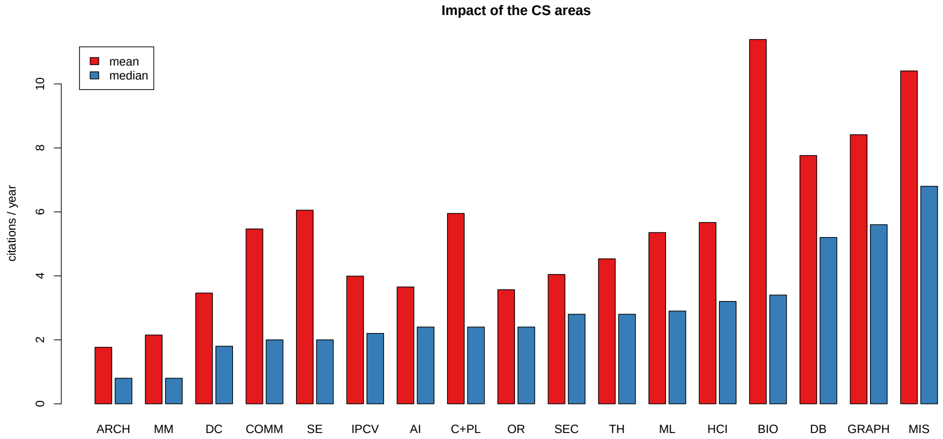

Figure 3: Citations per year, ordered by median.

3.5 H-indices

Figure 5 (and Table 8 in Appendix A) depicts the mean H-index for the sample of 20 researchers per area. Appendix C.7 presents the statistical significance results. None of the differences are statistically significant. Thus, the average CS researcher H-index (according to Google Scholar) is 14.9, independent of the research area.

4 Discussion and Conclusion

This research is a first attempt to measure the differences in publication and citation practices of some of the areas of Computer Science. In some aspects, the results of this research show information that researchers in the various areas already know, for example, the emphasis that some areas of CS have for conference publication. However, this research *quantifies* those intuitions and pre-empirical impressions.

The total productivity analysis shows that although the productivity of the areas range from 2.5 (MIS) to 7.8 (DC) papers per year, the only significant differences are among the extremes of the spectrum. Thus the total productivity of researchers in ARCH, COMM, DC, and IPCV is significantly higher than those working in MIS and OR. The productivity of the other areas are not significantly different. Thus CS departments and evaluation bodies should be mindful when comparing researchers in MIS and OR, to researchers in ARCH, COMM, DC, and IPCV.

Some evaluation criteria, especially criteria that apply to other disciplines besides CS put a larger emphasis on journal publication. CS departments that apply such emphasis on journal

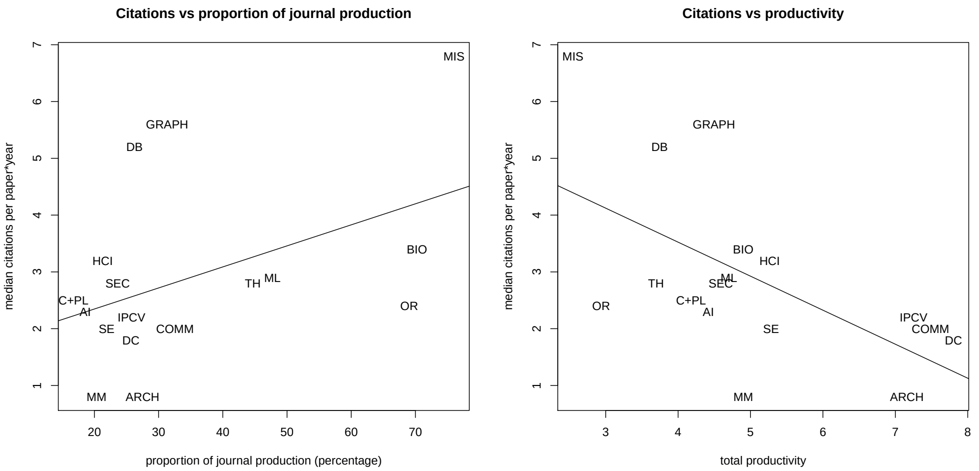

Figure 4: Citation rates vs proportion of journal publication and vs total productivity. A small jitter was added to C+PL and AI citation rates to distinguish the points.

publication must be mindful that BIO researchers in one group, and COMM and ML researchers in another have significantly different journal productivities than the other areas.

The differences in emphasis in publishing in journals or in conferences is more marked across the different areas. BIO, MIS, and OR are clearly journal oriented and significantly different than other areas. The areas or ML and TH are also significantly different than the most conference oriented areas.

Regarding citations, there are significant differences among MIS (by itself), BIO, DB, HCI and GRAPH (in another group), the other CS areas, and ARCH and MM. There is also an interesting negative correlation between productivity and citation rates, which goes beyond the influence of an area's emphasis on conference or journal publications.

Finally, there are no differences among the average H-index for researchers in any of the areas.

There are some other interesting findings:

- BIO and SEC were included as examples of new CS areas. BIO indeed has very different publication and citation patterns than the majority of the other CS areas. SEC's publication and citation patterns are not different than the majority.
- BIO, MIS, and OR are indeed less central CS areas, in the sense that a higher proportion of researchers in these areas are not in CS departments, but so is IPCV, to our surprise
- to the point that our sampling is able to reveal students preferences, the areas of ARCH, DB, HCI, and MM seems to be the most attractive to students, MIS seems to be the least.

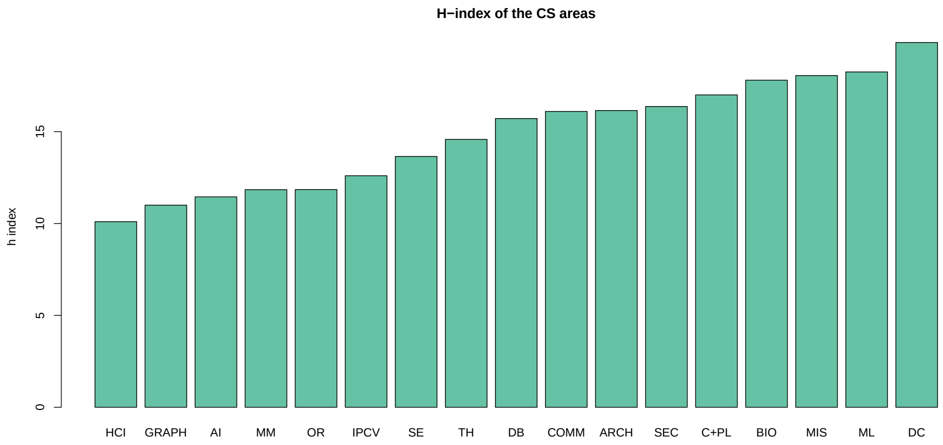

Figure 5: Mean H-index.

In spite of all of the above conclusions, this research should be taken only as a first step into that study. The conclusions herein are limited by at least four aspects that need further discussion. They are:

- the choice of areas;
- the definition of which venues are part of each area, and who are the researchers working in each area;
- the use of a sampling strategy to collect the data, especially publication data.
- the size of the research sample

Regarding the choice of areas, as we explained in Section 2.1, our decisions were based on a combination of goals: to use standard recognized areas, to include both “old” and “new” research areas, and to include areas that are not “central” to CS, in the sense that they are not always represented in CS departments. But ultimately, the definitions of what are the subareas of CS is *ad hoc*. Laender et al. [9] use 27 different areas while Biryukov and Dong [2] uses 14. Our definition is compatible with these two papers: all the areas we consider herein are included in [9] (which of course contains other areas, such as Robotics, Computer Education, and so on), but they group GRAPH and IPCV together. The fact that all the authors of that paper and most of the authors of this paper are Brazilians may indicate that these areas may reflect a national understanding of CS. Biryukov and Dong [2] defines only areas more “central” to CS, similar to ours, with some exceptions: (1) they group GRAPH and IPCV; (2) they include Data Mining explicitly with ML;

(3) they have Cryptography and Natural Language Processing plus Information Retrieval as areas by themselves; (4) they do not have Artificial Intelligence; and (5) their area of World Wide Web may or may not correspond to our MM.

As for the choice of venues and researchers for each area, the definition is *ad hoc* as well. Biryukov and Dong [2] define a set of important *conferences* for each area as the set of venues that represent that area, and define that a researcher works in the area if he or she co-authored at least one paper in these conferences. We were inspired by their methodology, however we wanted at least two further requirements: (1) our seed venues should not be only conferences but also journals, especially to include the areas which usually prefer journals as a publishing venue; and (2) we wanted an automatic way to safely expand this initial set of venues. Our idea is that it is not only the co-authors that publish in these central venues that represent the researchers in the area, and thus that is the justification for the iterative algorithm that expands the set of researchers we introduced in Section 2.3.

A first application of the aforementioned ideas generated a problem especially with the non-central CS areas. Let us exemplify the problem in the MIS area. The iterative algorithm given the seed (correctly) includes the journal *Management Science* as a venue for MIS because (1) that journal is in the DBLP database, (2) some of researchers that published in the MIS seed venues also published there, and (3) that was not true for any other of the areas. However, once that journal is included in the MIS set, any researcher that published one or more papers there would be considered a MIS researcher, and that is not correct — most of the authors in that journal are not MIS researchers, but business researchers. To address this problem, we include a new requirement which says that to be considered a researcher in a non-central CS area such as MIS, an author must have at least one paper published in the seed venues which, by construction, well represent the area.

For central areas in CS, this requirement probably eliminated from the population some of researchers that could be considered as members of that area's research community, but were not included because they never published in any of the seed venues. These researchers were not included and thus had no opportunity of being selected in the sampling procedure. On the other hand, no researcher that should not have been included in the population of researchers of an area was given the opportunity of being sampled.

The use of sampling as the strategy to collect the data, instead of using all the publication data available in some publication database is an uncommon choice. Most of the published work on bibliometric research in CS uses DBLP as the data source ([4, 2, 9, 5] to cite a few). We could not follow that approach. As it has been pointed out by [9, 13] DBLP has different coverage for different CS areas. If we had used DBLP as the data source one would not know if the differences in productivity were due to the different practices of researchers in different areas, or the differences in the DBLP coverages of these areas. Therefore, we used DBLP to define the populations and the set of researchers and venues in each CS area, but the final measurements of production were not based on DBLP, instead they were based on the papers listed on their personal webpages.

As a related result of the sample data collection, we evaluate DBLP's coverage for each CS area (Table 4). Notice that for DB and DC, DBLP *overestimates* the researcher production while it underestimates the researchers' production for areas such as BIO and IPCV.

Table 4: DBLP coverage for the different CS areas.

Area	DBLP coverage
AI	88%
ARCH	80%
BIO	57%
C+PL	101%
COMM	86%
DB	130%
DC	124%
GRAPH	76%
HCI	77%
IPCV	62%
MIS	87%
ML	83%
MM	92%
OR	75%
SE	82%
SEC	90%
TH	100%

Although our sampling method is advantageous over DBLP's different area coverage, it is worth noting that it also has downsides. The first, as we discussed in Section 2.4, is that the sample is not totally random. In a certain way, our procedure ends up giving preference to junior faculty in western universities, because they are the most likely researchers to keep an up-to-date publication list page. Given that junior faculty are the researchers that will most likely be evaluated using some of the metrics used in this paper, we believe that this bias has a small effect. However, faculty in non-western country universities should be careful when using the results in this paper, because they may not reflect their reality.

The second downside is that, since it is a costly procedure, only a small number of researchers were sampled, which is the fourth issue in this discussion. Thus the standard error associated with the measures is high and thus not all differences are statistically significant at the standard level of 95%. Thus some of our claims of non-significance may be revised if a larger sample is used. For example, we were surprised that there was no statistically significant differences between the group of higher productivity (ARCH, COMM, DC, IPCV) and the middle group, which contained all but (MIS and OR). A larger sample size may yet reveal that this difference is significant.

The procedure described in this paper is repeatable. The reader may choose a different set of areas and initial seeds to explore more specific questions. The costly step is defining the sample - finding which researchers have an up-to-date web page with his or her list of publications. A second costly step is collecting the H-index for the sample researchers. To help the researcher who wants to expand our results or explore other issues regarding publication and citation practices of the different areas of CS, we provide in www.ic.unicamp.br/~wainer/datasets/CSareas/ the data used in this research.

References

[1] S. D. J. Barbosa and C. S. de Souza. Are HCI researchers an endangered species in Brazil? *Interactions*, 18(3):69–71, 2011.

[2] M. Biryukov and C. Dong. Analysis of computer science communities based on DBLP. In *Research and Advanced Technology for Digital Libraries*, volume 6273 of *LNCS*, pages 228–235. Springer, 2010.

[3] C. Chen. Classification of scientific networks using aggregated journal-journal citation relations in the journal citation reports. *Journal of the American Society for Information Science and Technology*, 59(14):2296–2304, 2008.

[4] E. Elmacioglu and D. Lee. On six degrees of separation in dblp-db and more. *ACM SIGMOD Record*, 34(2):33–40, 2005.

[5] M. Franceschet. Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, 2011.

[6] A. Harzing. Publish or perish. available from http://www.harzing.com/pop.htm, 2007.

[7] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[8] J. Kapeller. Citation metrics: Serious drawbacks, perverse incentives, and strategic options for heterodox economics. *American Journal of Economics and Sociology*, 69(5):1376–1408, 2010.

[9] A. Laender, C. de Lucena, J. Maldonado, E. de Souza e Silva, and N. Ziviani. Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *ACM SIGCSE Bulletin*, 40(2):135–145, 2008.

[10] F. Lee, T. Grijalva, and C. Nowell. Ranking economics departments in a contested discipline: A bibliometric approach to quality equality between theoretically distinct subdisciplines. *American Journal of Economics and Sociology*, 69(5):1345–1375, 2010.

[11] B. Meyer, C. Choppy, J. Staunstrup, and J. van Leeuwen. Research evaluation for computer science. *Communications of the ACM*, 52(4):31–34, 2009.

[12] D. Patterson, L. Snyder, and J. Ullman. Evaluating computer scientists and engineers for promotion and tenure. Technical report, Computing Research Association, 1999. http://www.cra.org/resources/bp-memos/evaluating_computer_scientists_and_engineers_for_promotion_and_tenure/.

[13] F. Reitz and O. Hoffmann. An analysis of the evolving coverage of computer science sub-fields in the dblp digital library. *Research and Advanced Technology for Digital Libraries*, pages 216–227, 2010.

[14] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[15] J. Wainer, C. Billa, and S. Goldenstein. Invisible work in standard bibliometric evaluation of computer science. *Communications of the ACM*, 54(5):141–146, 2011.

A Complementary Tabular Results

Table 5: Number of researchers in specific areas. *Single* indicates the number of researchers that work only in that area. *Total* indicates the total number of researchers.

Area	**Single**	**Total**
AI	1698	2244
ARCH	3201	3662
BIO	7704	8406
C+PL	685	1001
COMM	3658	4395
DB	1359	1716
DC	635	1112
GRAPH	1732	2176
HCI	2815	3229
IPCV	5193	6826
MIS	1095	1175
ML	1871	2619
MM	2253	3623
OR	2766	3103
SE	1892	2278
SEC	1163	1527
TH	1126	1595

Table 6: Productivity of each area: conference papers, journal paper and total papers, per year.

Area	Conference		Journal		Total	
	mean	sd	mean	sd	mean	sd
AI	3.60	2.71	0.82	0.76	4.42	3.31
ARCH	5.19	5.01	1.97	2.30	7.16	6.57
BIO	1.46	2.16	3.44	3.64	4.90	4.38
C+PL	3.48	2.39	0.70	0.84	4.18	3.10
COMM	5.05	5.85	2.44	2.50	7.49	8.02
DB	2.76	2.46	0.98	1.12	3.74	3.32
DC	5.80	4.13	2.00	1.34	7.80	4.76
GRAPH	3.09	3.19	1.41	0.97	4.50	3.85
HCI	4.14	4.02	1.12	1.08	5.26	4.84
IPCV	5.38	6.20	1.87	1.52	7.25	7.14
MIS	0.61	1.15	1.94	0.83	2.55	1.28
ML	2.46	2.16	2.24	2.20	4.70	3.23
MM	3.90	2.78	1.00	0.92	4.90	3.39
OR	0.91	1.59	2.03	0.98	2.94	2.16
SE	4.13	3.25	1.16	1.50	5.28	3.94
SEC	3.51	3.76	1.08	1.27	4.59	4.94
TH	2.04	2.02	1.65	1.43	3.69	2.62

Table 7: Mean and median citations per year for papers in the areas.

Area	Mean	Median
AI	3.65	2.40
ARCH	1.77	0.80
BIO	11.39	3.40
C+PL	5.95	2.40
COMM	5.47	2.00
DB	7.76	5.20
DC	3.46	1.80
GRAPH	8.41	5.60
HCI	5.67	3.20
IPCV	3.99	2.20
MIS	10.41	6.80
ML	5.35	2.90
MM	2.15	0.80
OR	3.57	2.40
SE	6.05	2.00
SEC	4.04	2.80
TH	4.53	2.80

B Full name of the seed venues

AI AIJ: Artificial Intelligence Journal (Elsevier); JAIR: Journal on Artificial Intelligence Research; AAAI: AAAI Conference on Artificial Intelligence; IJCAI: International Joint Conference on Artificial Intelligence

BIO BMC Bioinf: BMC Bioinformatics; Bioinformatics: Bioinformatics (Oxford Journals); JCB: Journal on Computational Biology (Mary Ann Liebert Publoshers); RECOMB: Conference on Research on Computational Molecular Biology; TCBB: IEEE/ACM Transactions on Computational Biology and Bioinformatics

C+PL OOPSLA: Conference on Object-Oriented Programming, Systems, Languages & Applications; POPL: ACM Symposium on Principles of Programming Languages; PLDI: Conference on Programming Language Design and Implementation; TOPLAS: ACM Transactions on Programming Languages and Systems; CGO: International Symposium on Code Generation and Optimization

COMM TON: IEEE/ACM Transactions on Networking; TCOM: IEEE Transactions on Commu-

Table 8: Mean H-index for the areas.

Area	Mean	Median
AI	11.45	8.00
ARCH	16.15	11.00
BIO	17.80	12.00
C+PL	17.00	15.50
COMM	16.10	10.50
DB	15.71	12.50
DC	19.85	14.00
GRAPH	11.00	9.50
HCI	10.10	5.50
IPCV	12.60	11.50
MIS	18.05	17.00
ML	18.25	10.00
MM	11.84	12.00
OR	11.85	9.00
SE	13.65	14.00
SEC	16.37	11.00
TH	14.58	10.00

nications; Mobicom: ACM International Conference on Mobile Computing and Networking; SIGComm: Conference of the ACM Special Interest Group on Data Communication; Infocom: IEEE International Conference on Computer Communications

ARCH ISCA: ACM IEEE International Symposium on Computer Architecture; MICRO: IEEE/ACM International Symposium on Microarchitecture; DAC: Design Automation Conference; ASPLOS: Architectural Support for Programming Languages and Operating Systems; TCAD: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems; SC: International Conference for High Performance Computing, Networking, Storage, and Analysis

GRAPH TOG: ACM Transactions on Graphics; CGA: IEEE Computer Graphics and Applications; TVCG: IEEE Transactions on Visualization and Computer Graphics; SIGGRAPH: ACM SIGGRAPH International Conference

DB TODS: ACM Transactions on Database Systems; VLDB: Very Large Data Bases Conference; Sigmod: ACM SIGMOD Conference

DC TPDS: IEEE Transactions on Parallel and Distributed Systems; JPDC: Journal of Parallel and Distributed Computing (Elsevier); ICDCS: International Conference on Distributed

Computing Systems; ICPP: International Conference on Parallel Processing

HCI TOCHI: ACM Transactions on Computer-Human Interaction; IJMMS: International Journal of Human-Computer Studies (Elsevier); UMUAI: User Modeling and User-Adapted Interaction - The Journal of Personalization Research; CHI: ACM Conference on Human Factors in Computing Systems; CSCW: ACM Conference on Computer Supported Cooperative Work

IPCV International Journal of Computer Vision (Springer); TIP: IEEE Transactions on Image Processing; CVPR: IEEE Conference Computer Vision and Pattern Recognition; ICIP: IEEE International Conference on Image Processing

ML JMLR: Journal of Machine Learning Research; ML: Machine Learning (Springer); NECO: Neural Computation (MIT); NIPS: Annual Conference on Neural Information Processing Systems; ICML: International Conference on Machine Learning

MIS ISR: Information Systems Research (INFORMS); MANSCI: Management Science (INFORMS); JMIS: Journal of Management Information Systems; EJIS: European Journal of Information Systems (Palgrave Macmillian); MISQ: MIS Quarterly

MM MMS: Multimedia Systems (Springer); TMM: IEEE Transactions on Multimedia; IEEEMM: IEEE MultiMedia; MM: ACM Multimedia (Conference); ICMCS: International Conference on Multimedia Computing and Systems

OR Math Prog: Mathematical Programming (Springer); SIOPT: SIAM Journal on Optimization; C&OR: Computers & Operations Research (Elsevier); Disc Appl Math: Discrete Applied Mathematics (Elsevier)

SEC TISSEC: ACM Transactions on Information and System Security; JCS: Journal of Computer Security; IEEESP: IEEE Security and Privacy; SP: IEEE Symposium on Security and Privacy; USS: USENIX Security Symposium; CCS: ACM Conference on Computer and Communications Security

SE TSE: IEEE Transactions on Software Engineering; TOSEM: ACM Transactions on Software Engineering and Methodology; ICSE: International Conference on Software Engineering; TACAS: International Conference on Tools and Algorithms for the Construction and Analysis of Systems; ESE: Empirical Software Engineering (Springer)

TH JACM: Journal of the ACM; SICOMP: SIAM Journal on Computing; STOC: ACM Symposium on Theory of Computing; FOCS: IEEE Annual Symposium on Foundations of Computer Science; SODA: ACM-SIAM Symposium on Discrete Algorithms

C Statistical Significance

C.1 Journal productivity

P-values for the pairwise comparisons between *journal* productivity for all areas. This table should be interpreted in the following way. The column c and row r intersection indicates the p-value of the pairwise comparison (with Holm correction) between the set of journal productivity measures for areas c and r. If the p-value is less than 0.05 than one can accept that the difference in journal productivity between areas c and r is statistically significant with 95% confidence. The entries with p-value less than 0.05 are in bold in the table below.

	AI	ARCH	BIO	C+PL	COMM	DB	DC	GRAPH	HCI	IPCV	MIS	ML	MM	OR	SE	SEC
ARCH	0.69															
BIO	**0.00**	0.06														
C+PL	1.00	0.33	**0.00**													
COMM	**0.01**	1.00	1.00	**0.00**												
DB	1.00	1.00	**0.00**	1.00	**0.02**											
DC	0.41	1.00	**0.05**	0.18	1.00	0.71										
GRAPH	1.00	1.00	**0.00**	1.00	1.00	1.00	1.00									
HCI	1.00	1.00	**0.00**	1.00	0.20	1.00	1.00	1.00								
IPCV	1.00	1.00	**0.03**	0.68	1.00	1.00	1.00	1.00	1.00							
MIS	1.00	1.00	0.07	0.54	1.00	1.00	1.00	1.00	1.00	1.00						
ML	0.10	1.00	0.53	**0.04**	1.00	0.18	1.00	1.00	0.89	1.00	1.00					
MM	1.00	1.00	**0.00**	1.00	0.07	1.00	1.00	1.00	1.00	1.00	1.00	0.41				
OR	0.44	1.00	0.09	0.20	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00			
SE	1.00	1.00	**0.00**	1.00	0.29	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
SEC	1.00	1.00	**0.00**	1.00	0.15	1.00	1.00	1.00	1.00	1.00	1.00	0.71	1.00	1.00	1.00	
TH	1.00	1.00	**0.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

C.2 Conference productivity

P-values for the pairwise comparisons between *conference* productivity for all areas. See explanation above.

	AI	ARCH	BIO	C+PL	COMM	DB	DC	GRAPH	HCI	IPCV	MIS	ML	MM	OR	SE	SEC
ARCH	1.00															
BIO	1.00	**0.00**														
C+PL	1.00	1.00	1.00													
COMM	1.00	1.00	**0.01**	1.00												
DB	1.00	0.38	1.00	1.00	0.53											
DC	1.00	1.00	**0.00**	0.76	1.00	**0.02**										
GRAPH	1.00	1.00	1.00	1.00	1.00	1.00	0.21									
HCI	1.00	1.00	0.32	1.00	1.00	1.00	1.00	1.00								
IPCV	1.00	1.00	**0.00**	1.00	1.00	0.21	1.00	1.00	1.00							
MIS	0.15	**0.00**	1.00	0.24	**0.00**	1.00	**0.00**	0.79	**0.02**	**0.00**						
ML	1.00	0.27	1.00	1.00	0.38	1.00	**0.02**	1.00	1.00	0.15	1.00					
MM	1.00	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.05**	1.00				
OR	0.27	**0.00**	1.00	0.42	**0.00**	1.00	**0.00**	1.00	**0.04**	**0.00**	1.00	1.00	0.10			
SE	1.00	1.00	0.36	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.03**	1.00	1.00	**0.05**		
SEC	1.00	1.00	1.00	1.00	1.00	1.00	0.82	1.00	1.00	1.00	0.22	1.00	1.00	0.39	1.00	
TH	1.00	0.06	1.00	1.00	0.09	1.00	**0.00**	1.00	1.00	**0.03**	1.00	1.00	1.00	1.00	1.00	1.00

C.3 Total productivity

P-values for the pairwise comparisons between *total* productivity for all areas. See explanation above.

	AI	ARCH	BIO	C+PL	COMM	DB	DC	GRAPH	HCI	IPCV	MIS	ML	MM	OR	SE	SEC
ARCH	1.00															
BIO	1.00	1.00														
C+PL	1.00	1.00	1.00													
COMM	0.79	1.00	1.00	0.46												
DB	1.00	0.18	1.00	1.00	**0.05**											
DC	0.29	1.00	1.00	0.16	1.00	**0.01**										
GRAPH	1.00	1.00	1.00	1.00	1.00	1.00	0.38									
HCI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00								
IPCV	1.00	1.00	1.00	0.99	1.00	0.16	1.00	1.00	1.00							
MIS	1.00	**0.02**	1.00	1.00	**0.00**	1.00	**0.00**	1.00	1.00	**0.01**						
ML	1.00	1.00	1.00	1.00	1.00	1.00	0.66	1.00	1.00	1.00	1.00					
MM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
OR	1.00	**0.03**	1.00	1.00	**0.01**	1.00	**0.00**	1.00	1.00	**0.03**	1.00	1.00	1.00			
SE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
SEC	1.00	1.00	1.00	1.00	1.00	1.00	0.49	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
TH	1.00	0.35	1.00	1.00	0.12	1.00	**0.03**	1.00	1.00	0.29	1.00	1.00	1.00	1.00	1.00	1.00

C.4 Proportion of journals

P-values for the pairwise comparisons between *proportion of journal* to total productivity for all areas. See explanation above.

	AI	ARCH	BIO	C+PL	COMM	DB	DC	GRAPH	HCI	IPCV	MIS	ML	MM	OR	SE	SEC
ARCH	1.00															
BIO	**0.00**	**0.00**														
C+PL	1.00	1.00	**0.00**													
COMM	0.28	1.00	**0.00**	**0.01**												
DB	1.00	1.00	**0.00**	0.22	1.00											
DC	1.00	1.00	**0.00**	0.67	1.00	1.00										
GRAPH	**0.00**	**0.03**	**0.00**	**0.00**	1.00	0.51	0.27									
HCI	1.00	1.00	**0.00**	1.00	1.00	1.00	1.00	**0.02**								
IPCV	1.00	1.00	**0.00**	0.38	1.00	1.00	1.00	0.86	1.00							
MIS	**0.00**	**0.00**	1.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**						
ML	**0.00**	**0.02**	**0.00**	**0.00**	1.00	0.35	0.19	1.00	**0.02**	0.65	**0.00**					
MM	1.00	1.00	**0.00**	1.00	0.65	1.00	1.00	**0.01**	1.00	1.00	**0.00**	**0.00**				
OR	**0.00**	**0.00**	1.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.00	**0.00**	**0.00**			
SE	1.00	1.00	**0.00**	1.00	0.19	1.00	1.00	**0.00**	1.00	1.00	**0.00**	**0.00**	1.00	**0.00**		
SEC	1.00	1.00	**0.00**	1.00	1.00	1.00	1.00	**0.03**	1.00	1.00	**0.00**	**0.02**	1.00	**0.00**	1.00	
TH	**0.00**	**0.00**	**0.00**	**0.00**	1.00	**0.04**	**0.02**	1.00	**0.00**	0.10	**0.00**	1.00	**0.00**	**0.00**	**0.00**	**0.00**

C.5 Citations

P-values of the pairwise comparisons between the median citations per year for all areas.

	AI	ARCH	BIO	C+PL	COMM	DB	DC	GRAPH	HCI	IPCV	MIS	ML	MM	OR	SE	SEC
ARCH	**0.00**															
BIO	1.00	**0.00**														
C+PL	1.00	**0.02**	1.00													
COMM	1.00	**0.00**	1.00	1.00												
DB	**0.05**	**0.00**	1.00	0.67	0.32											
DC	1.00	0.07	0.11	1.00	1.00	**0.00**										
GRAPH	**0.00**	**0.00**	1.00	**0.05**	**0.02**	1.00	**0.00**									
HCI	1.00	**0.00**	1.00	1.00	1.00	1.00	0.39	0.71								
IPCV	1.00	**0.02**	0.94	1.00	1.00	**0.03**	1.00	**0.00**	1.00							
MIS	**0.00**	**0.00**	1.00	**0.01**	**0.00**	1.00	**0.00**	1.00	0.11	**0.00**						
ML	1.00	**0.00**	1.00	1.00	1.00	1.00	0.17	0.58	1.00	1.00	0.10					
MM	**0.01**	1.00	**0.00**	0.08	**0.01**	**0.00**	0.24	**0.00**	**0.00**	0.07	**0.00**	**0.00**				
OR	1.00	**0.02**	0.67	1.00	1.00	**0.02**	1.00	**0.00**	1.00	1.00	**0.00**	1.00	0.06			
SE	1.00	**0.00**	1.00	1.00	1.00	0.25	1.00	**0.02**	1.00	1.00	**0.00**	1.00	**0.01**	1.00		
SEC	1.00	**0.00**	1.00	1.00	1.00	0.11	1.00	**0.01**	1.00	1.00	**0.00**	1.00	**0.01**	1.00	1.00	
TH	1.00	**0.00**	1.00	1.00	1.00	0.64	1.00	**0.05**	1.00	1.00	**0.01**	1.00	**0.00**	1.00	1.00	1.00

C.6 Linear regression of citations

The output of the analysis of the coefficients for the regression of the median citation rate against total production, and proportion of journal publication.

```
lm(formula = cit.median ~ total.prod + prop.journal)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1656 -0.5826 -0.2218  0.5625  2.5021

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72857    1.69465   2.790   0.0145 *
total.prod  -0.49230    0.24355  -2.021   0.0628 .
prop.journal 0.01859    0.01969   0.944   0.3612
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1.348 on 14 degrees of freedom
Multiple R-squared: 0.3804,Adjusted R-squared: 0.2919
F-statistic: 4.298 on 2 and 14 DF,  p-value: 0.03506
```

C.7 H-index

P-values of the pairwise comparisons between the mean H-index per year for all areas.

	AI	ARCH	BIO	C+PL	COMM	DB	DC	GRAPH	HCI	IPCV	MIS	ML	MM	OR	SE	SEC
ARCH	1.00															
BIO	1.00	1.00														
C+PL	1.00	1.00	1.00													
COMM	1.00	1.00	1.00	1.00												
DB	1.00	1.00	1.00	1.00	1.00											
DC	1.00	1.00	1.00	1.00	1.00	1.00										
GRAPH	1.00	1.00	1.00	1.00	1.00	1.00	1.00									
HCI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00								
IPCV	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00							
MIS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00						
ML	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00					
MM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
OR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00			
SE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
SEC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
TH	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

D Final venues for each area

This section lists the final venues for each area. The “ j/” prefix indicates that the venue is a journal; the “ c/” prefix, a conference. The abbreviations are the ones used in DBLP. The DBLP page for a journal with abbreviation `jjj` is at `http://www.informatik.uni-trier.de/~ley/db/journals/jjj/`. For a conference with abbreviation `ccc`, the DBLP page can be found at `http://www.informatik.uni-trier.de/~ley/db/conf/ccc/`.

AI j/ai, j/jair, j/jar, c/aaai, c/ijcai, j/logcom, j/ki, j/japll, c/comma, j/synthese, c/aips, c/aisc, c/rr, c/ruleml, c/atal, c/lpnmr, j/constraints, c/iclp, c/icail, j/esi, c/cia, j/sLogica, c/cp, j/aamas, c/promas, j/tplp, c/ecsqaru, j/aim, j/aicom, c/mates, c/jelia, c/dlog, c/kr, c/aimsa, c/ecai

ARCH c/iscas, c/micro, c/dac, c/asplos, j/tcad, c/sc, c/3dic, c/ats, c/apccas, j/todaes, c/dsd, j/ejes, c/socc, c/asap, c/ahs, c/dft, j/trets, c/recosoc, j/mj, c/vts, j/micro, c/cases, j/jetc, c/samos, j/jolpe, j/dt, c/fccm, j/iet-cdt, c/isqed, j/ijes, c/fdl, c/iccad, c/aspdac, c/glvlsi, c/date, c/sbcci, c/estimedia, c/fpl, c/fpt, j/vlsi, c/delta, c/ddecs, j/et, c/nocs, c/vlsi, c/isca, c/reconfig, c/fpga, c/ets, c/arc, j/tvlsi, c/vlsid, c/ersa, c/isvlsi, c/islped, c/ispd, c/iolts, c/codes, j/amcs, j/integration, j/jcsc, c/iccd, c/rsp, c/patmos

BIO j/bmcbi, j/bioinformatics, j/jcb, c/recomb, j/tcbb, c/bibm, c/bibe, c/ismb, c/apbc, c/bird, j/ijdmb, c/bcb, c/f-ic, j/tcsb, j/nar, j/candc, c/wabi, c/psb, j/jcisd, c/gcb, c/ijcbs, j/jamia, c/dils, j/jcamd, j/ijbra, j/jib, c/isbra, j/jbcb, c/cmsb, j/bib, c/mie, c/bicob, c/prib, j/pcs, c/cibcb, c/biocomp, j/almob, j/jbi, j/ijcbdd, c/medinfo, j/ejbsb

C+PL c/oopsla, c/popl, c/pldi, j/toplas, c/cgo, c/aplas, j/cl, c/esop, c/cc, c/icfp, j/jfp

COMM j/ton, j/tcom, c/mobicom, c/sigcomm, c/infocom, j/twc, j/bell, c/imc, c/networking, c/wiopt, c/ipom, j/ccr, c/pam, j/jsac, c/ciss, j/ett, j/ijscn, c/broadnets, c/isita, j/osn, c/valuetools, c/conext, c/comsware

DB j/tods, c/vldb, c/sigmod, c/edbt, c/cidr, j/debu, j/sigmod, j/dpd, c/ssd, j/vldb, j/pvldb

DC j/tpds, j/jpdc, c/icdcs, c/icpp, c/ISCApdcs, c/ispdc, j/tjs

GRAPH j/tog, j/cga, j/tvcg, c/siggraph, c/apvis, c/si3d, c/rt, c/apgv, c/tpcg, j/tap, c/egpgv, c/npar, c/sca, c/graphite, c/vissym, c/mig, j/cgf, c/simvis, j/cagd, c/egve

HCI j/tochi, j/ijmms, j/umuai, c/chi, c/cscw, c/avi, c/tamodia, c/mhci, j/interactions, j/puc, j/ijwbc, c/ACMdis, c/eics, c/ihm, c/interact, c/pdc, c/uist, c/ozchi, c/nordichi, c/tabletop, c/sigdoc, c/mc, j/ctw, c/candc, j/pervasive, j/icom, j/cscw, j/cscl, c/pervasive, c/tei, c/ah, c/huc, c/persuasive, j/jcmc, c/w4a, c/wmte, j/iwc, j/ais, j/uais, c/acmidc, c/group, c/usab, c/bcshci, c/assets, j/chb, j/ijhci, j/behaviourIT, j/hf

IPCV j/ijcv, j/tip, c/cvpr, c/icip, c/ibpria, c/scia, c/gbrpr, c/icisp, c/ipas, c/emmcvpr, c/bmvc, c/caip, j/jdi, c/3dpvt, j/pami, c/miar, c/cimaging, c/amdo, c/ipmi, c/crv, c/fimh, c/isbi, j/tmi, c/icb, c/ciarp, j/siamis, c/scalespace, j/mva, c/icvgip, j/tgrs, c/miccai, j/jmiv, c/eccv, c/iccv, c/wbir, c/hvei, j/mia, c/dagm

MIS j/isr, j/mansci, j/jmis, j/ejis, j/misq, j/isj, j/ism, j/iam, j/jitech, c/sigcpr, j/irmj, j/bise, j/jais, j/msom, j/isem, j/dss, j/mktsci, j/ijeis, j/da, j/db, j/misqe

ML j/ml, j/jmlr, j/neco, c/nips, c/icml, j/tnn, c/uai, c/esann, c/pkdd, j/bc, c/alt, c/icgi, j/jcns, c/ecml, j/nn

MM j/mms, j/tmm, j/ieeemm, c/mm, c/icmcs, c/ltconf, c/trecvid, c/elpub, j/jdim, j/ejasmp, c/amr, c/civr, c/delfi, c/mmm, j/aml, j/amm, c/ism, c/pcm, c/nossdav, c/ichl, c/mir, c/samt, j/jmm, j/ijdet, c/wiamis, j/mta, c/3pgcic, c/semco, c/ircdl, c/iscslp, c/doceng, j/ijsc, j/tomccap

OR j/mp, j/siamjo, j/cor, j/dam, j/anor, j/eor,j/ol, j/dmgt, c/ifip5-7, j/jgo, c/or, j/arscom, j/jors, j/disopt, j/apjor, j/accs, j/candie, j/order, j/dm, j/networks, j/endm, c/colognetwente, j/orl, j/scheduling, j/informs, j/4or, j/gc, j/transci, j/heuristics

SE j/tse, j/tosem, c/icse, c/tacas, j/ese, c/issta, c/aswec, j/ijseke, c/fase, j/sopr, j/sttt, c/oss, j/re, c/issre, c/icfem, c/msr, c/re, c/ecsa, j/sigsoft, c/aose, j/sqj, c/jiisic, c/splc, c/euromicro, j/infsof, j/isse, c/wosp, c/spin, c/quatic, c/forte, c/pts, c/ecmdafa, c/csmr, c/qosa, c/kbse, c/models, c/vamos, j/sosym, c/sigsoft, j/fmsd, c/asm, c/wicsa, j/software, c/apsec, j/iee, c/refsq, j/fac, c/xpu, c/icsm, c/icsr, c/icst, c/iceccs, c/rcis, c/wcre, c/formats, c/enase, c/ispw, c/profes, c/csee, c/edoc, j/ase, c/iwpc, c/fm, c/cibse, j/smr, c/wer, c/cbse, c/esem

SEC j/tissec, j/jcs, j/ieeesp, c/sp, c/uss, c/ccs, c/europki, c/raid, c/icisc, c/iwsec, c/iciss, c/icics, j/ijisec, c/wpes, c/pet, c/acisp, c/ctrsa, c/policy, j/virology, c/indocrypt, c/isw, c/acsac, c/trustbus, c/provsec, c/esorics, c/fc, c/dbsec, c/sec, c/cans, c/acns, c/ifip11-9, c/ifip1-7, c/ndss, c/sacmat, c/csfw, c/pkc, c/pairing

THEO j/jacm, j/siamcomp, c/stoc, c/focs, c/soda, c/compgeom, c/waoa, c/esa, c/fsttcs, c/icalp, c/cccg, j/eccc, j/cpc, j/comgeo, j/cc, c/wine, j/algorithmica, j/dcg, j/combinatorica, c/swat, j/apal, j/jcss, c/coco, j/jsyml, c/wads, j/mst, c/sigecom, j/rsa, c/latin, c/isaac, j/im, c/stacs, c/approx, c/sagt, j/talg