CACM Gender Computing

Introduction

Five years after the publication of the extensive bibliometric overview of publishing by women in ACM conference papers (Cohoon, Nigai, and Kaye 2011) it seemed time to see whether the positive trends spotted in that paper persisted and to widen the analysis.

In this paper we give a bibliometric analysis of the dynamics in the relationships between gender and *productivity* (measured as number of publications), *collaboration* (measured by coauthorships), and *scientific impact* (measured by citations and h-index) in the field of computing. Our results are based on DBLP, the bibliographic database for the field of computing (Ley 2005). The aim of DBLP is to be complete for this field, at least for the more recent period, which results in a wide variety of different types of publications. For this reason, we restricted the analysis to the period 1990-2015 and to the roughly 100 top rated conferences and journals in computing (as established by (Computing Research and Education Association of Australasia (CORE) 2014)).

Our dataset includes the top rated ACM conferences, but also others like WWW, IJCAI and the top IEEE conferences, and top journals, mostly from ACM, IEEE, Elsevier and Springer. We are predominantly interested in *growth rates*, both absolutely (e.g. absolute share of all computing publications by women), and relatively (e.g., does the impact of publications grows faster for women than for men?).

Main results

- Women still do worse than men
- On all measures women grow faster than men
- Hardly any difference between journals and conferences
- Now half of all papers has at least 1 female author.
- there are quite some differences between fields inside computing

Prior research

todo

Data and methods

Data

Our main data source is the DBLP XML dump which can be downloaded from http://dblp.uni-trier.de/xml/. A useful feature for our research is that DBLP normalizes author names and, when available, uses the full first name in the normalized name (Ley 2005).

We view the data as a bipartite affiliation network with a set of articles and a set of authors as nodes, and one relation between the nodes of these two sets: authorship. Authors and articles each have one property, their gender and their field, respectively. In the analysis to come, we mostly work with the edges in this network: the authorships, a pair consisting of an article and an author.

** See Thomas scriptie**

Inferring gender from names

Almost every study which relates gender with other other variables has to deal with the fact that gender is not explicitly available in public data. Within DBLP there is very little intrinsic information about the authors: their first and last name and rarely an affiliation and a URL to a homepage. To infer gender we built an ensemble classifier based on existing software which infers gender from first names and from portrait photographs. We did an extensive evaluation on two datasets with ground truth available: 600 authors from DBLP for which gender and a portrait photograph were obtained by students; and the almost 1 million persons with a page in the English Wikipedia with a gender. The best combination of first name and image classifiers received an F2 score of .90 on both datasets, which is higher than any separate classifier achieves.

For this study we used only the first names. The F2 scores are then marginally higher for women than for men.

** Detailed results in appendix: See Thomas scriptie**

Results

In this section we study the changes in the relation between gender and productivity, collaboration and impact. We start with an overview of the data.

DBLP restricted to the A* rated conferences and journals contains 756K authorships. Of these 85.5% are assigned a gender. The dataset on which all analyses are done contains all publications which appeared in the interval 1990-2015, borders included, with all authorships without gender removed. This set contains

559K authorships from 220K articles with in total 172K different authors. Two third of all authorships are in conferences, and 64% of all articles appeared in conferences. The mean number of authors is higher at conferences than at journals (3.9 vs 3.5).

In the dataset, 22% of the authors is female. Of the conference and journals authorships, 20 and 19% is female, respectively.

Productivity

Figure 1 shows the growth in the number of authorships per year per gender. The number of publications are in log scale. Female authorship is growing faster than male authorship with respective regression coefficients of 0.043 and 0.031. This means that if we extrapolate these two growth rates that in 2060 both genders produce and equal amount of papers.

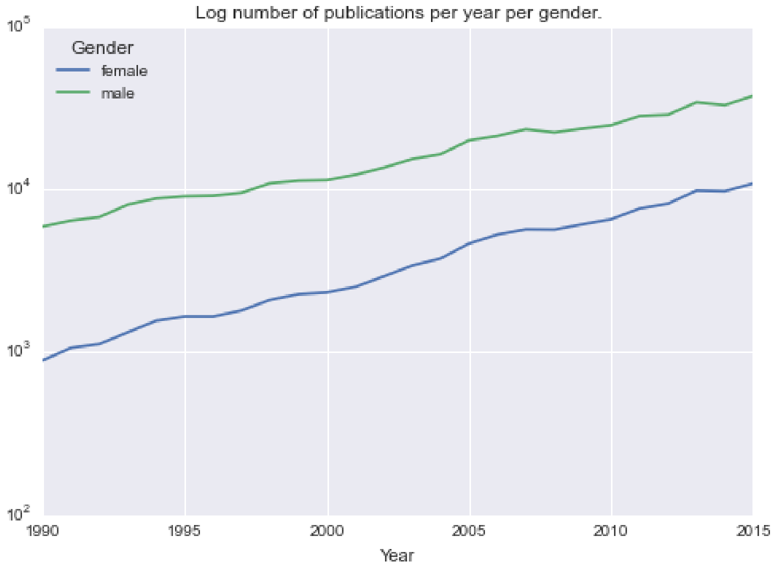

Figure 1: Log number of publications per year per gender

If we measure the influence of female authors differently we see a much faster growth, see Figure 2. The number of publications with at least one female author increases linearly with 1 percentage point per year, both for conferences and journals. In 2015, 47 and 46% of the papers in conferences and journals had at least one female author. The percentages are higher for conference papers as these have on average more coauthors.

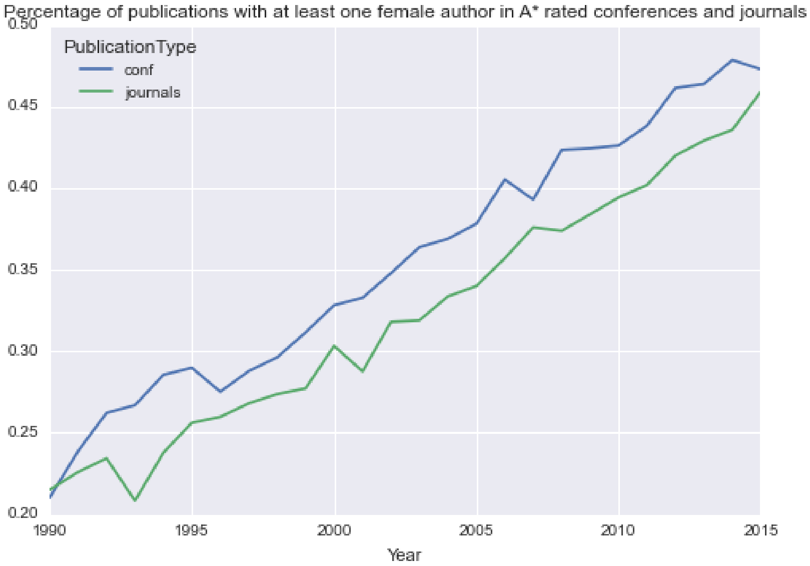

Figure 2:

The most natural unit of counting productivity is the authorship, a pair consisting of an article and an author. We can count absolute authorships or, as in done in (Sugimoto et al. 2013), relative authorships, in which we account for the number of authors by dividing authorship by the number of all authors of the article (whether or not their gender is inferred). Figure 3 shows that the two ways of counting show the same picture: a linear growth of 0.4 percentage point in female authorships per year. As before the growth rate is the same for conferences and journals, but the percentage is 2 percentage points higher for conferences (not shown in the figure).

Earlier studies showed that women are less likely to appear as first or last author of a paper (Sugimoto et al. 2013,@10.1371/journal.pone.0066212). Figure ?? shows that this is the case too in the field of computing, but that at least the ratios are rising at roughly the same speed. First and last authorships show an interesting pattern. Women are 2 percentage point less often last author and 2 percentage point more often first author than the 22% (in 2015) they have of all authorships. For last authorship this most probably reflects the fact the women are underrepresented in the higher faculty positions. In computing, often the PhD advisor of the main (and first) author takes the last authorship. If we assume that PhD students are overrepresented as first authors, than the finding is in line with (Cohoon, Nigai, and Kaye 2011) which find that female Ph.D.s publish more than male ones.

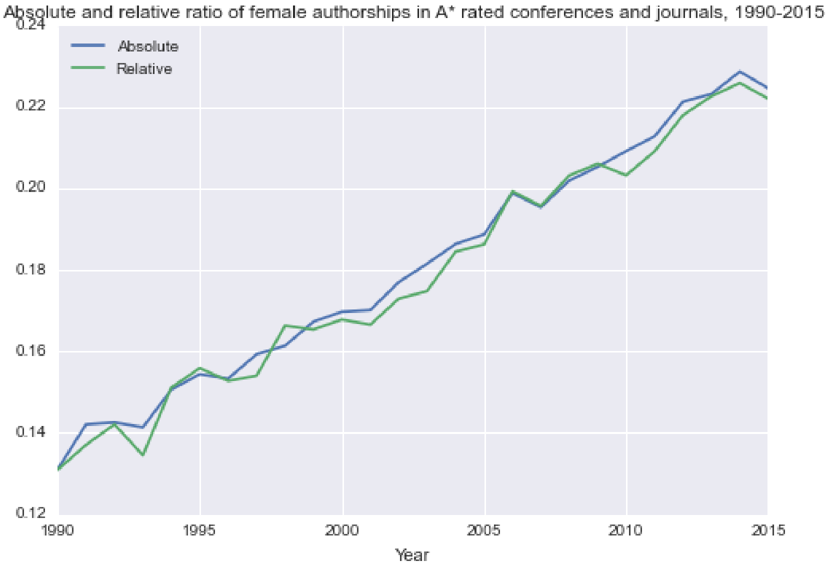

Figure 3:

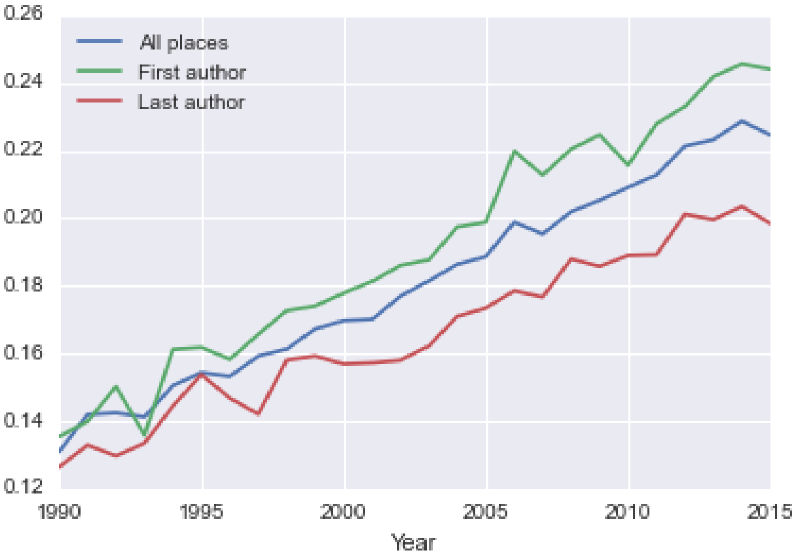

Figure 4: Ratio of absolute female authorships, in any place, in first place and in last place, in A* rated conferences and journals.

Figure 5 shows the growthrate of female authorships for 5 different subfields within computing, together with the regression lines. We restricted the data to fields with at least 10K female authorships in total. The table below gives the increase in percentage point per year for the different fields. For each field we list the conferences included in that field in the CORE listing. Note that we also included journals, but they are not listed here.

If we compare the rates in these subfields with the overal growth of 0.4 percentage point, we notice that AI performs on average, web and data base and datamining perform above average and programming languages below average. Field 806 has the lowest growth but that is a field with an already high proportion of female authorships. These results are in line with those of (Cohoon, Nigai, and Kaye 2011), who find that *'the greatest gender differences were evident in conferences focusing on Human Factors, Languages, Algorithms, and Performance, in decreasing order.'*

Field	**Increase per year**	**Conferences**
805	0.59	IEEE INFOCOM,PERCOM,PERVASIVE,PODC,UbiComp,WWW
804	0.53	CRYPTO,DCC,EuroCrypt,ICDE,ICDM,PODS,SIGKDD,SIGMOD,VLDB,WS
801	0.37	AAAI,AAMAS,ACL,COLT,FOGA,ICAPS,ICCV,ICML,IEEE InfoVis,IJCAI,IJ
803	0.30	ACMMM,ASPLOS,CCS,HPCA,ICFP,ICSE,ISCA,OOPSLA,OSDI,PLDI,POPI
806	0.23	CHI,ICIS,ISWC,JCDL ,SIGIR

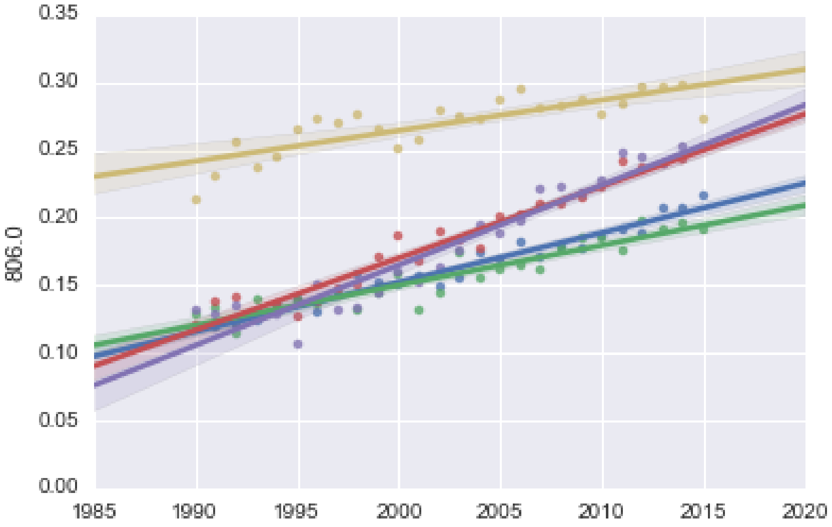

Figure 5:

Cooperation

We now study how the cooperation between the sexes has changed over the years. For that we compute two conditional probabilities from the data: given that a paper has an author of one sex, and at least two authors, what is the probability that this paper is coauthored with a female? Over all years these probabilities are 0.14 and 0.26, for given female and male authors, respectively. Both probabilities increase over the years, with 0.30 and 0.42 percentage point per year, respectively. Figure 6 shows the ratios per year.

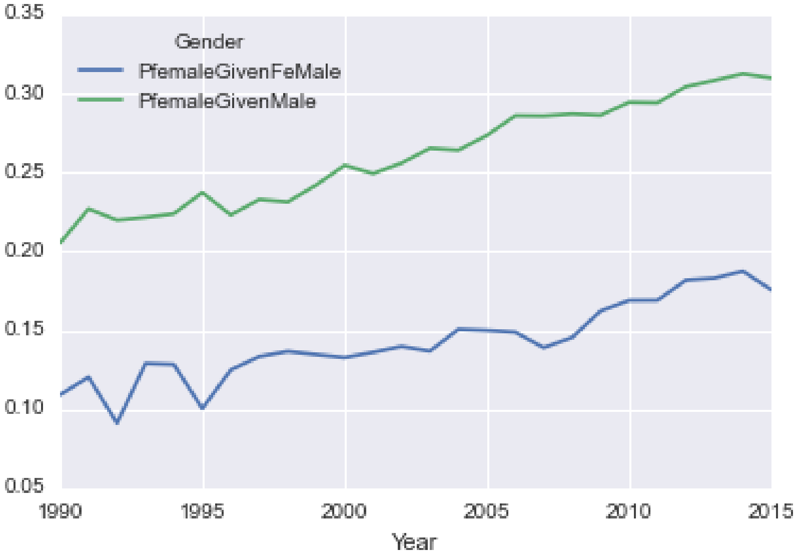

Figure 6: Conditional probabilities of a paper having a female coauthor given that it has a male or give that it has a female author.

Impact

We measure change in impact in two ways: by using the impact measures computed by Google Scholar, and by adding the citation counts from Google Scholar for each article to the list of articles in DBLP and following the same procedure as in the productivity analysis.

We start with the measures by Google Scholar.

Impact measures calculated by Google Scholar

We compare 3 measures of impact of publications: number of citations, H-index and I10-index, all computed by Google Scholar. We matched authors from DBLP to Google Scholar and if we obtained a unique hit, then we scraped the information from the authors page at Google Scholar. We sorted the DBLP authors inversely by the number of publications in DBLP. After several months of scraping we crawled 51K authors of which 70% was male, 15% female and of 15% we could not assign gender. Google provides two measurement points for the three impact measures: at 2011 and the current (mid 2016) value. We computed the percentage growth for males and females. Thus we removed authors which had no measurement in 2011, leaving 34K males and 8K females. We also removed the top 250 authors with most citations (the maximum was 226K). In this top, 9% was female.

On all three impact measures we see significant differences in growth. As these distributions are heavily skewed, we used the Mann Whitney test for significance testing. The table below contains the median growth percentages for the three measures.

	citations	**h index**	**i10 index**
female	41.0	17.0	27.0
male	48.0	20.0	30.0

Figure 7 shows the number of authors with a certain percentage growth in the number of citations. We see that the shapes of the curves of both genders are similar, but the female growth is consistently below that of the male. The figures for the two index measures are similar.

Citation analysis per publication

We crawled citation counts for publications in DBLP from Google Scholar using the following procedure: we sorted the authors in DBLP on number of publications, and downloaded their list of publications from Google Scholar. We then matched these with the publications of that author from DBLP, normalizing the title and controlling for year and type of publication. In this way positive citation counts were found for 112K unique articles. We removed articles without any counts from the analysis, and removed authorships without gender, leading to 317K authorships of which 19.5% by female authors.

Female authors are slightly overrepresented in those publications with few (between 1 and 10) citations (21.5%). For our further analysis we removed all articles with fewer than 10 (N=93K) and more than 250 (N=18K) citations. On this set 18.8% of all authorships was by a female author. This set contains 75K papers with an average of 3.1 coauthors. In total there are 58K male and 15K

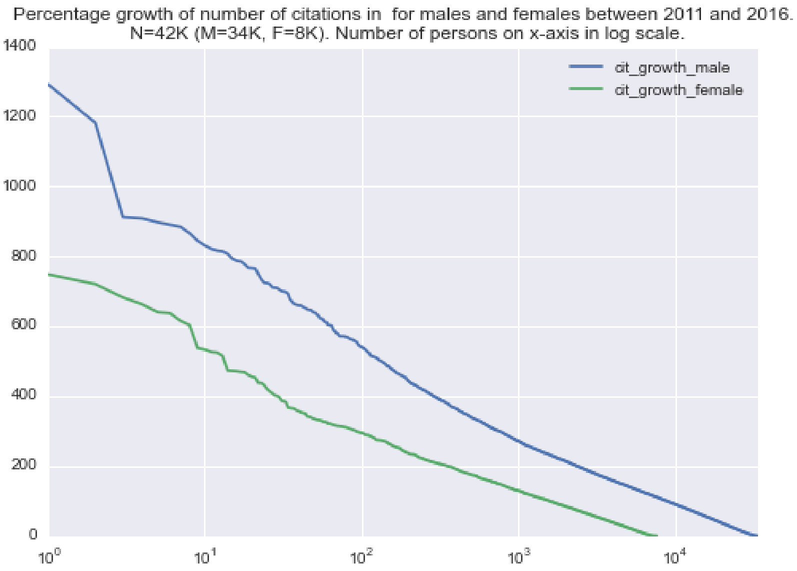

Figure 7: Percentage growth of number of citations in for males and females between 2011 and 2016. N=42K (M=34K, F=8K). Number of persons on x-axis in log scale.

female authors. Even though we removed publications with very low and very high number of citations, the number of citations per author is highly skewed. The table below lists the mean and median number of citations per gender. Figure 8 shows the median number of citations per author per year grouped by gender. The difference between the two distributions per year is not significant (Wilcoxon sign test: p=0.02).

Gender	Mean number of citations	Median number of citations
Male	164	63
Female	141	55

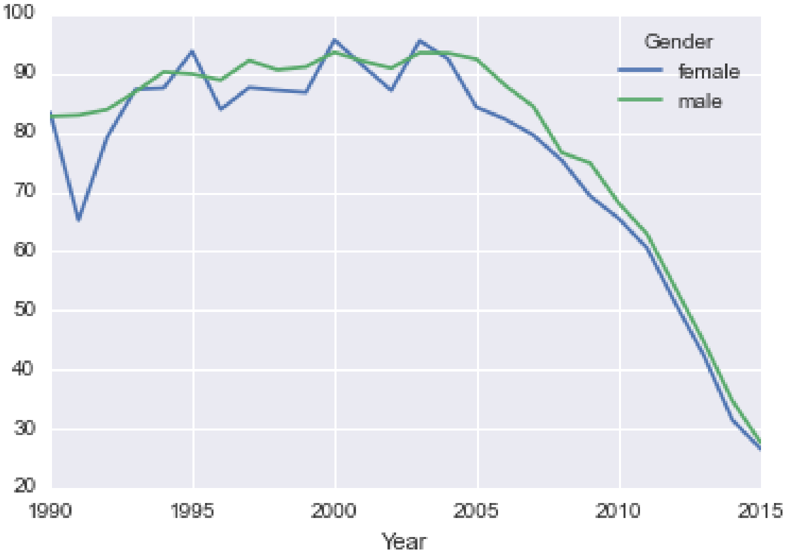

Figure 8: Median number of citations per author per year grouped by gender.

References

Cohoon, J McGrath, Sergey Nigai, and Joseph Jofish Kaye. 2011. “Gender and Computing Conference Papers.” *Communications of the ACM* 54 (8). ACM: 72–80.

Computing Research and Education Association of Australasia (CORE). 2014. “CORE Rankings Portal.”

Ley, Michael. 2005. "DBLP Computer Science Bibliography." University of Trier.

Sugimoto, Cassidy R, Vincent Lariviere, CQ Ni, Yves Gingras, Blaise Cronin, and others. 2013. "Global Gender Disparities in Science." *Nature* 504 (7479). Macmillan Publishers Ltd., London, England: 211–13.