How Inclusive are We?
An Analysis of Gender Diversity in Database Venues

Angela Bonifati
Lyon 1 University & Liris CNRS
France
angela.bonifati@univ-lyon1.fr

Michael J. Mior
Rochester Institute of Technology
Rochester, NY
mmior@cs.rit.edu

Felix Naumann
Hasso Plattner Institute
University of Potsdam, Germany
felix.naumann@hpi.de

Nele Sina Noack
Hasso Plattner Institute
University of Potsdam, Germany
nele.noack@student.hpi.de

ABSTRACT

ACM SIGMOD, VLDB and other database organizations have committed to fostering an inclusive and diverse community, as do many other scientific organizations. Recently, different measures have been taken to advance these goals, especially for underrepresented groups. One possible measure is double-blind reviewing, which aims to hide gender, ethnicity, and other properties of the authors.

We report the preliminary results of a gender diversity analysis of publications of the database community across several peer-reviewed venues, and also compare women's authorship percentages in both single-blind and double-blind venues along the years. We also obtained a cross comparison of the obtained results in data management with other relevant areas in Computer Science.

1. INTRODUCTION

Increasingly, the computer science and database community are recognizing the importance of actively increasing diversity, in particular gender diversity among researchers, or removing impediments to the advancement of underrepresented researchers in the field. For instance, ACM SIGMOD and VLDB together started an initiative "to create an inclusive and diverse database community with zero tolerance for abuse, discrimination, or harassment", and the D&I in DB initiative coordinates such efforts across the data management community[1].

One opportunity to increase diversity might be double-blind reviewing, hiding the authors names and thus effectively hiding their gender from the reviewers. While there might be further signals about the gender of the author(s), for instance in their writing style or the topic of the paper, author names are the most direct indicators of gender to reviewers and readers. Moreover, gender lookup using names has also been adopted in recent work on the authorship of Computer Science contrasted with other fields of study [10]. Only in an (albeit large) minority of cases the gender is not unambiguously revealed by the first name even if the reviewer does not personally know the author and their gender. Other methods, such as targeted surveys in our community or image processing on photos of personal homepages, could be used to address these ambiguous cases. These methods go beyond the scope of our work and are the subject of future investigation.

In this study, we analyze and compare the participation of women in papers at various top-level conference and journals. To this end, we make use of a commercial service to assign gender based on first names for many languages. While we realize that gender is not a binary concept distinguishing women and men, we do not have the means to identify any more fine-grained designations based on the given data, which matches that which reviewers and readers usually have at their disposition. Next, we have downloaded and prepared reference data from DBLP. With our dataset, we are able to compare the evolution of such diversity across the years and compare the diversity across venues, some of which perform double-blind reviewing. Our analysis considers only accepted papers; we do not report about the diversity of rejections due to lack of data.

Our preliminary findings show that there is an overall growth of the number of accepted papers authored by women in major database conferences, with some slight differences. We also examined how the data management field stands with respect to other fields such as HCI, AI, Algorithms, Network-

[1] `https://dbdni.github.io/`

ing, and Operating Systems. In this landscape, the differences might also be due to the gender-composition of the researchers in the respective fields. Finally, we could not observe a tangible difference between single-blind and double-blind reviewing for the data concerning the SIGMOD conference. The analysis of the submission data could be enlightening in that case.

The following Section 2 discusses related work. Then, Section 3 introduces both our approach to identify the gender of authors and the considered publication datasets. Section 4 is the core of this empirical paper, presenting our analytical findings. Finally, we conclude with an outlook on possible further analyses in Section 5.

2. RELATED WORK

Snodgrass provides an excellent survey of literature analyzing the effects single- vs. double-blind reviewing [6], which we do not repeat in our empirical work here. Many studies from different research fields do mention gender fairness as a goal of double-blind reviewing. However, the cited results are often inconclusive: some report a significant bias, others do not observe this. Snodgrass concludes [6]: "These studies show that revealing author identity, specifically the gender of the author, can sometimes have an effect on acceptance rates."

In the database research field, the SIGMOD conference is a particularly interesting venue to analyze: until the year 2000 it employed single-blind reviewing before switching to double-blind reviewing in 2001. Apart from gender bias, the original impetus for this change, and for double-blind reviewing in general, is to avoid any bias of reviewers to more favorably review and to more readily accept papers by well-known, prolific authors, and to thus let the content speak for itself. We are not the first to analyze the effects of this change of reviewing policy. Madden and DeWitt identified "prolific" authors and their success rate at SIGMOD and VLDB conferences from 1995 until 2005 [3]. They conclude that "double-blind reviewing has had essentially no impact on the publication rates of more senior researchers in the database field". Tung performed a similar study on the same data, concluding "that there are indications that double-blind reviewing does have an impact in terms of papers accepted for famous people in SIGMOD" [9]. However, neither of the two works addresses gender diversity.

Tomkins et al. also analyzed the impact of double-blind reviewing using data from a single computer science conference edition: WSDM'17. Here, some reviewers had access to author information while others did not [7]. In their study they also analyze the "Matilda effect", in which "publications from male authors are associated with greater scientific quality, in particular if the topic is male-typed" [2]. Tomkins et al. [7] found no statistically significant impact on bidding and reviewing both for papers with a woman as first author and for papers with a majority of women as authors. They do perform a meta-analysis across seven studies, which, put together, show a statistically significant negative bias for these papers.

Other analyses of bibliometric data from DBLP-DB have been carried out in the past, e.g., to study the collaboration network in our community [1]. That study shows that there is a power law on the frequency of publications and presents other statistics, such as the number of co-authors per scholar. They do not discuss the impact of gender in this kind of analysis.

3. PREPARING, SELECTING, AND ANALYZING DATA

In this section, we explain how we selected and preprocessed the data used in our analysis. We also discuss how we carried out our assessment. Our entire analysis is reproducible and the source code along with additional results are publicly available[2].

3.1 Defining paper gender

In this paper, we focus on gender analysis of bibliographic data in the data management field. While other analyses could be done by considering diversity of the writing style, paper topics, or other factors, we do not consider them here. We focus on authorship information for a paper and define three different categories of gender when associating it to a paper.

- A paper whose first author is a woman (FAW)
- A paper whose last author is a woman (LAW)
- A paper with any author being a woman (AAW)

Clearly, papers that fit the first two definitions also fit the last definition, but not vice versa. These definitions are sufficient to let us take an initial dip into the analysis and study the trends of woman authorship in our community. We distinguish the three aforementioned definitions in our analysis and show and cross-compare the corresponding results.

[2] `https://github.com/HPI-Information-Systems/GenderAnalysis/`

Alternative definitions are clearly possible to study the data under different perspectives and by considering other dimensions in addition to gender. For instance, one can think of analyzing bibliographic data by looking at other diversity criteria, which are equally important, such as race, ethnicity, country of origin, culture, affiliation, (academic) age, etc. Although these criteria are applicable to our corpus, we do not regard them here.

3.2 First-name analysis

Automatically deriving gender from first names is known to be a very difficult problem [5]. Some rules of thumb might apply. For instance, knowing the gender of first names in case of familiarity with the language of the country of origin of that author makes sense as an applicable rule. However, in some spoken languages, there might exist ambiguity in the gender of first names. For instance, Andrea is typically a woman-identified first name in Germany, whereas it is exclusively a men-identified first name in Italy. The same first name is sporadically used for men in Germany for people being Italian immigrants. In these cases, the country of origin of the authors could help us disambiguate the gender of the authors. While it would be possible to use country of origin in the DBLP data in order to help disambiguate the names, this affiliation country data is quite sparse ($< 30\%$) and we decided not to use it in this first analysis.

From the list of publications, we infer the authors' full names and split them into first, (middle), and last names. For obtaining the genders of the first names, we use Gender API[3], a commercial online platform to determine gender by first names. In the first step, we use the list of first names to look up the gender. If the first name is abbreviated, we look to the middle name(s). For a given first name, Gender API provides the predicted binary gender along with an estimated accuracy and the number of samples of that name held in their database. We use the predicted gender if the accuracy is higher than 50% percent. Otherwise, we label the first name concerned with 'neutral'. There are also some names for which Gender API does not provide any result. We label these names as well as fully abbreviated ones with 'unknown'.

To not under-represent either men or women, we consider the gender of all unknown and gender-neutral names to randomly be either man or woman, based on the overall gender distribution in the portion of the data where the predicted gender is more certain. We are aware that this binary assignment does not respect all genders and that the extending the observed women/men distribution to all other names might introduce some bias. Furthermore, the name someone is given at birth may not necessarily be one that matches their gender identity. However, as our goal is to assess potential bias among reviewers, we expect the gender commonly perceived to be associated with a particular author's name to be a sufficient starting point for this analysis. We also tried alternative distributions, e.g., unknown gender data considered all men or all women, and observed that the overall trends of accepted papers for women did not change and no further insights could be gleaned from the obtained results.

3.3 Venue selection

Our data is taken from the DBLP computer science bibliography[4]. We downloaded the entire proceedings data available in DBLP for a selection of popular database research and other CS venues and collected all authorship information. Our analysis includes ACM SIGMOD, VLDB, ICDE, EDBT and CIDR conferences. Notice that among these, only SIGMOD is double-blind, while the remaining ones are single-blind. For the data concerning VLDB, we combined the conference data (VLDB) with the data from Proceedings of VLDB (PVLDB), the latter being the replacing journal starting from 2008. We label the combination as VLDB.

For comparison, we planned to also include other top database journals, such as VLDB Journal and ACM Transactions on Database Systems (TODS). Due to the low absolute number of papers appearing in TODS, we decided to dismiss it in the presentation of the results. Finally, we include a lower-ranked conference (DASFAA) and a lower-ranked journal (DKE) to allow a comparison between higher and lower ranked venues. Table 1 lists for each venue the years for which we gathered data, and the overall number of papers for that duration.

Venue	**Years**	**# pubs**	**# authors**
CIDR	2003 – 2020	476	1,173
DASFAA	1989 – 2020	1,939	4,220
DKE	1985 – 2020	1,719	3,438
EDBT	1988 – 2020	1,552	3,307
ICDE	1984 – 2020	4,743	8,046
SIGMOD	1975 – 2020	4,065	6,959
(P)VLDB	1975 – 2020	5,198	8,621
VLDBJ	1992 – 2020	907	1,996

Table 1: Captured years and number of papers for each conference

[3]`https://gender-api.com/`

[4]`https://dblp.org/`

Furthermore, we cross-compare the data in our community with neighboring communities in computer science. For that purpose, we regarded CS-Rankings[5], considered the data of selected fields, and chose the corresponding conferences listed there, as reported in Table 2.

Field	Venues	avg. authors
AI	AAAI, IJCAI	3.10
Algorithms	FOCS, SODA, STOC	2.44
Databases	SIGMOD, VLDB, ICDE, PODS	3.44
HCI	CHI, UIST, UbiComp, Pervasive, IMWUT	3.93
Networking	SIGCOMM, NSDI	4.20
Operating Systems	OSDI, SOSP, EuroSys, FAST, USENIX ATC	4.34

Table 2: Venues listed for other fields

4. DIVERSITY RESULTS

In this section, we report the results of our analysis concerning (i) papers authored by women accepted in the data management community across the years and venues listed in Table 1, and (ii) trends of accepted papers in neighboring communities in computer science for the fields and conferences listed in Table 2. Across all figures, we report a 3-year moving average percentage of papers following in each category.

Figures 1, 2 and 3 show this average for the three categories of the first (FAW), last (LAW), and any author (AAW) having a woman-identified name across all years in which that venue published papers. By looking at the results, we can observe the following:

- CIDR shows the lowest diversity across all categories, but, being a single-track conference and being a biannual event until recently, the overall number of papers is lower compared to other venues leading to low significance of our analysis. Moreover, the conference was limiting the number of papers submitted by the same author (to 1 or 2 depending on the years) and focusing solely on systems, vision, and prototype papers.

- For SIGMOD we created two regression lines: one up to 2000 for its single-blind process, and one from 2001 onward to reflect its double-blind process. We did not observe a remarkable difference in the percentage of accepted papers by women after shifting to a double-blind review policy. However, we cannot draw a conclusion on this aspect, as this would require inspecting more data (including the submission data).

- By examining the FAW results, we can observe a higher percentage of papers accepted in DASFAA, which could suggest that women as first authors are more successful in this conference. However, this trend is less prominent in the LAW and AAW results for DASFAA. A similar trend can be observed for DKE with peaks in the period 1995-2000 for all three percentages.

As a disclaimer for the results reported above, we let the reader notice that the outcome of our analysis should be taken with some caution. Indeed, the presence of authors with unknown genders for which we did infer the gender and the fact that we collapsed the entire proceedings into one bulk piece of data (without distinction between long and short papers with different respective acceptance rates) might lead to some confounding factors. As such, our analysis is preliminary and can certainly be improved in future work.

Finally, Figure 4 shows the results for the percentage of papers authored by women (FAW) aggregated per CS field. We chose five additional research fields as reported in Table 2.

From these results, we can observe that the HCI field sees the highest percentages of papers by women across the years, whereas the Operating Systems field is lowest. We can also see that, at least recently, the database field is faring somewhat better than the remaining fields. Nevertheless, these results should be taken with a grain of salt since they also depend on the gender composition of the various fields. In particular, we point out that the information about gender composition of the different fields is missing at present, as also highlighted in recent work on the dynamics of gender bias [4]. Once this information will be available, it can help interpret better the above results.

5. CONCLUSION

In this paper, we have focused on the gender impact on authorship in the data management area. We started from the assumption that women are an underrepresented group in computing [10, 11]. This assumption has been confirmed by the results of our study.

Our analysis was of course only a preliminary step towards many and more detailed analyses. For

[5] `http://csrankings.org/`

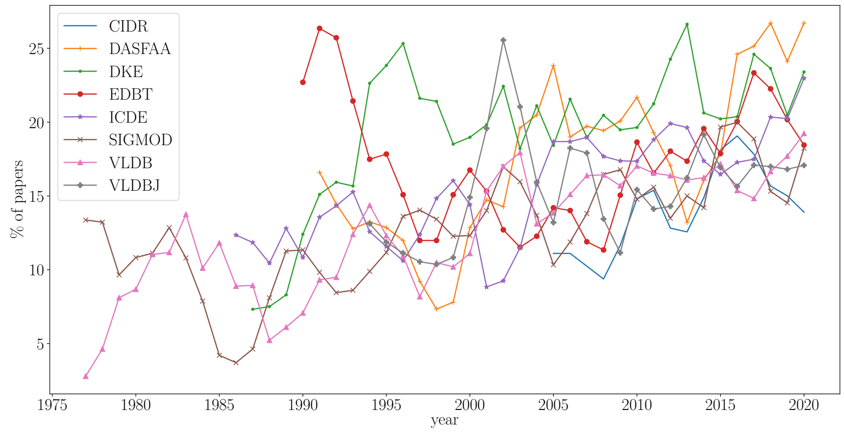

Figure 1: First author woman (FAW) percentages by year (3-year moving average)

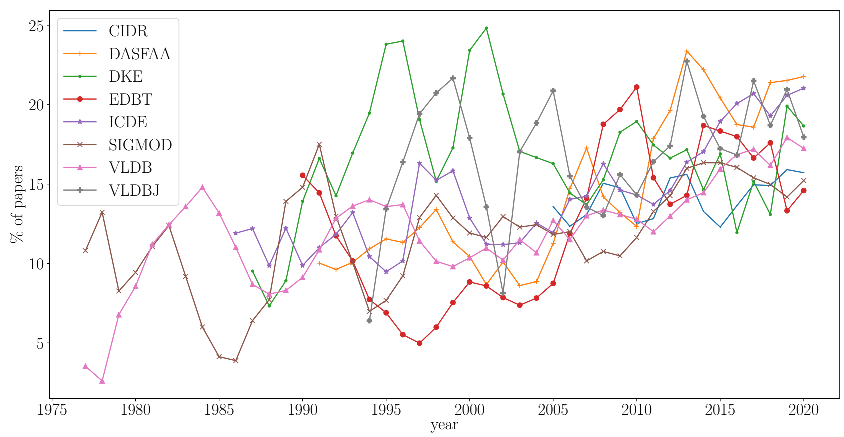

Figure 2: Last author woman (LAW) percentages by year (3-year moving average)

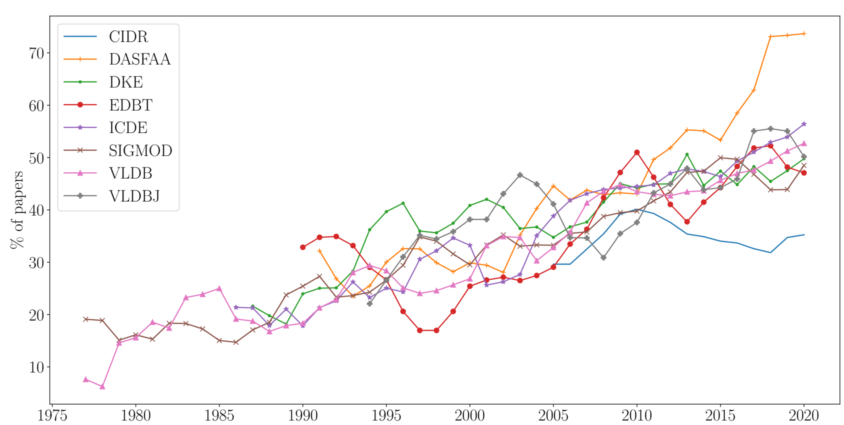

Figure 3: Any author woman (AAW) percentages by year (3-year moving average)

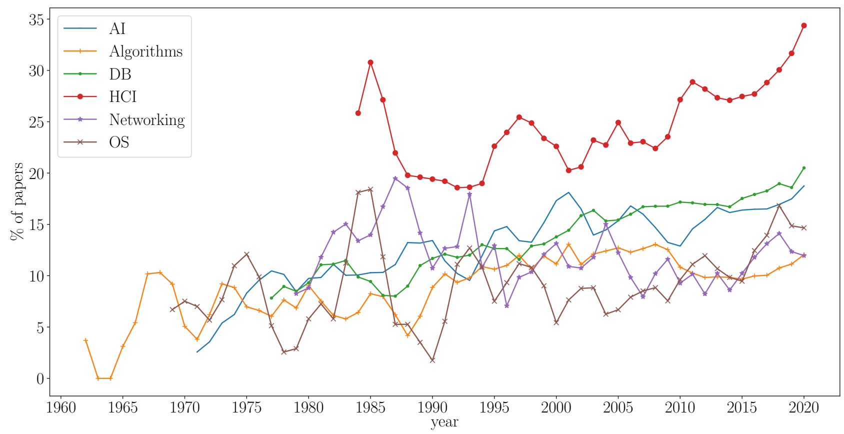

Figure 4: First author woman (FAW) percentages across fields in CS (3-year moving average)

instance, with affiliation data, the same statistics could be broken down by region, by country or by individual affiliation. Gender assessment using names can become more accurate by leveraging manual annotations and targeted surveys within our community or by image processing starting from website pictures, even if the latter has other limitations, such as solely considering gender as binary, the inherent noise of the available data, etc. Also, while the overall trends show an increase in diversity, it would be interesting to compare gender with the academic age to validate the hypothesis that this increase is mostly due to junior women entering the field.

An even more insightful analysis could be performed not only on accepted papers, as we do here, but including also data about submitted papers to the various venues. The latter would be more difficult, as it requires accessing sensitive data, such as the submission data and reviews for conferences and journals in our field. Moreover, this analysis would be applicable to one conference and one edition of the conference only, as it has been done for instance for the ICLR conference [8].

Acknowledgements

We would like to thank the colleagues who helped annotate first names: Mazhar Hameed, Hazar Harmouch, Lan Jiang, Ioannis Koumarelas, Nitisha Jain, Chao Zhang, Meihui Zhang, and Kyuseok Shim. We also thank Renée Miller for starting this discussion at VLDB 2020 and Arun Kumar for giving us inspiring and very helpful feedback.

6. REFERENCES

[1] Ergin Elmacioglu and Dongwon Lee. On six degrees of separation in DBLP-DB and more. *SIGMOD Record*, 34(2):33–40, 2005.

[2] Silvia Knobloch-Westerwick, Carroll J. Glynn, and Michael Huge. The Matilda effect in science communication: An experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, 35(5):603–625, 2013.

[3] Samuel Madden and David DeWitt. Impact of double-blind reviewing on SIGMOD publication rates. *SIGMOD Record*, 35(2):29–32, June 2006.

[4] Thomas J. Misa. Dynamics of gender bias in computing. *Communications of the ACM*, 64(6):76–83, 2021.

[5] Lucía Santamaría and Helena Mihaljević. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, page 4:e156, 2018.

[6] Richard Snodgrass. Single- versus double-blind reviewing: An analysis of the literature. *SIGMOD Record*, 35(3):8–21, 2006.

[7] Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences (PNAS)*, 114(48):12708–12713, 2017.

[8] David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of OpenReview: A critical analysis of the machine learning conference review process. *CoRR*, abs/2010.05137, 2020.

[9] Anthony K. H. Tung. Impact of double blind reviewing on SIGMOD publication: A more detail analysis. *SIGMOD Record*, 35(3):6–7, September 2006.

[10] Lucy Lu Wang, Gabriel Stanovsky, Luca Weihs, and Oren Etzioni. Gender trends in computer science authorship. *Communications of the ACM*, 64(3):78–84, 2021.

[11] Telle Whitney and Valerie Taylor. Increasing women and underrepresented minorities in computing: The landscape and what you can do. *Computer*, 51(10):24–31, 2018.