Gender Representation in Brazilian Computer Science Conferences

Natália Dal Pizzol, Eduardo Dos Santos Barbosa, Soraia Raupp Musse

[1]School of Technology - Pontifical Catholic University of Rio Grande do Sul
Av. Ipiranga, 6681 – 90619-900 – Porto Alegre – RS – Brazil

natalia.pizzol,eduardo.santos94@edu.pucrs.br, soraia.musse@pucrs.br

Abstract. *This study presents an automated bibliometric analysis of 6569 research papers published in thirteen Brazilian Computer Science Society (SBC) conferences from 1999 to 2021. Our primary goal was to gather data to understand the gender representation in publications in the field of Computer Science. We applied a systematic assignment of gender to 23.573 listed papers authorships, finding that the gender gap for women is significant, with female authors being under-represented in all years of the study.*

1. Introduction

Although female figures have played a crucial role in the development and advancement of Computer Science (CS), the field has become predominantly male [Wang et al. 2021]. This gender disparity is considered a global phenomenon, and one way that it presents itself in Brazil is in the low rates of enrollment for female students in Computer Science degrees. The 2019 Statistics on College Education Report, published by the Brazilian Computer Society (SBC)[1], shows that women represent only 11.51% of incoming CS students.

Despite increased discussions and concerns about the inclusion of women in Computer Science communities, evidence has not yet been obtained that can clearly state how this disparity in the gender proportion impacts research publications in Brazilian conferences. Research publications are the main instrument to disseminate scientific knowledge [Holman et al. 2018], and in the case of Brazilian CS researchers, works are often published in conference proceedings [Albertini et al. 2019].

According to the Brazilian Computer Society's list of events[2], many conferences in Computer Science have been created in the last 20-30 years in Brazil. The exact number of conferences active in 2022 is particularly difficult to estimate if we consider regional or local workshops. However, if we select conferences supported by the Brazilian Computer Society, for which many papers are available in public data sets, we can achieve the objective of finding Brazilian conferences to evaluate the diversity in gender representation of published authors.

The goal of this work is to present an analysis of the diversity and representation of gender in Brazilian conferences over time by conducting a bibliometric study based on systematic assignment of gender. This article is organized as follows: Section 2 presents works related to the topic of diversity and female representation in Computer Science.

[1]https://www.sbc.org.br/documentos-da-sbc/category/133-estatisticas
[2]https://www.sbc.org.br/eventos/eventos-realizados

Section 3 outlines the proposed methodology for the study, including data collection, gender inference and implementation. Section 4 details the results obtained in this research. Finally, Section 5 presents our final considerations.

2. Related Work

The lack of diversity and absence of a variety of different voices in Science, Technology, Engineering and Mathematics (STEM), has been a frequent topic of study in recent years. Haines et al. [Haines et al. 2020] have investigated the influence of the researcher's gender on their research topic by analyzing birdsong literature. In animal behavior studies, birdsong was studied as a primarily male trait, however, they have found that women working on the field are more likely to be the authors of female birdsong articles. The significant contributions that women have made to the field of female birdsong studies suggest that diversity in academia can foster new scientific ideas, maximizing the value and quality of research.

Nevertheless, even with studies showing the importance of diversity in science, gender disparities persist. In 2013, Larivière et al. [Larivière et al. 2013] conducted a global multidisciplinary study on scientific publications indexed in Thomson Reuters Web of Science databases between 2008 and 2012. They found that, globally, female authorship accounted for fewer than 30% of publications, and that articles with women in dominant author positions tend to receive fewer citations. Their results showed how prevalent gender inequality is in STEM.

In the field of Computer Science, several studies focused on mapping global gender representation. Wang et al. [Wang et al. 2021] analyzed gender trends in Computer Science literature with a sample of 11.8M articles published from 1970 to 2019 that were indexed in the Semantic Scholar literature corpus. Their findings show that, although the proportion of female authors is increasing, there is still a notable gender gap in the academic authorship of CS research.

The overall under-representation of women in CS research is well-established, but other studies have been conducted to scan the distribution of this inequality across CS subfields. Cheon et al. [Cheong et al. 2021] have analyzed nine subfields, and found that female authors are outnumbered in each one of them, and that the fields of Artificial Intelligence, Information Security, Computer Vision, Machine Learning, and Systems Architecture represent the less gender-diverse areas, with an approximate 5:1 ratio of male to female authors.

Ribeiro et al. [Ribeiro et al. 2019] have conducted a quantitative analysis of SBC members, investigating their gender, location, type of membership, and areas of interest in CS. They found that 77,71% of SBC members are male and 21,67% are female. Regarding areas of interest in CS, they found that there is a bigger proportion of women interested in the areas of Information Systems, Multimedia and Hypermedia Systems, Collaborative Systems, Informatics in Education, and Human Computer Interaction, while the least favored areas by women are Computational Architecture and High Performance Processing, Distributed Systems, Integrated Circuit Design, Computer Networks and Distributed Systems, and Algorithms, Combinatorics and Optimization.

Arruda et al. [Arruda et al. 2009] have analyzed 886 publications from 2000 to 2006 authored by Brazilian researchers. Their work classifies researchers into CS sub-

fields, and suggests that female scientists tend to concentrate in the areas of Artificial Intelligence, Collaborative Systems, Computer in Education, and Human-Computer Interfaces. This research, however, shows some limitations since they were working with a small sample of authors.

Although several contributions have been made to the study of gender representation in CS, fewer articles focus on analyzing Brazil specifically. In order to contribute to this investigation, the main research question in our work concerns gender representation in Brazilian Computer Science conferences. Next section presents the applied methodology.

3. Methodology

The primary goal of this study is to assess the gender diversity and representation of authors who contribute to advancing computer science research in Brazil. To that end, we analyzed publications from thirteen SBC conferences, as listed in Table 1. Those 13 conferences were selected from the total of 31 events listed in the SBC Open Library of publications[3]. There were three main criteria for the selection of the conferences in this work:

- the conference must be sponsored by the Brazilian Computer Society[4]. This criteria aims to avoid local or regional and small workshops.
- the conference should have had at least 20 editions. This criteria was used in order to select more consolidated conferences in Brazil.
- the conference should be indexed in either the Scopus or DBLP databases. Finally, this criteria was chosen to identify papers that have more scientific visibility.

More details on the data collection and gender inference processes are provided in the following subsections.

3.1. Data Collection and Processing

There are several available databases to consult computer science publications, some of the most commonly used are DBLP, Scopus, and Web of Science. This study was conducted using DBLP[5], due to its high index of unique articles [Cavacini 2014], programmer-friendly API, and free service. We used DBLP API's query for venues, which returns a list of every indexed publication for a determined conference, journal, etc. We applied the names of the selected conferences, presented in Table 1, as query parameters. This DBLP API query returned the associated metadata for each publication, allowing for the retrieval of title, author and coauthors, year of publication, type of publication (such as conference or workshop papers), DOI (Digital Object Identifier) and venue information.

The next step was to sort and filter the publication information. We chose to exclude publications that did not contain DOI information in the associated metadata, as the DOI is one of the most reliable identifiers of a publication. The publication's DOI ensured that no articles in our list were repeated, and increased the transparency of the research since all works could be referenced and located.

[3] https://sol.sbc.org.br/index.php/anais/confs
[4] sbc.org.br
[5] https://dblp.org/

Table 1. SBC conferences analyzed in the study.

Conference Name	Reference	Editions*
Brazilian Symposium on Human Factors in Computational Systems	IHC	20
Symposium on Computer Architecture and High Performance Computing	SBAC-PAD	33
Brazilian Symposium on Databases	SBBD	36
Brazilian Symposium on Integrated Circuits and Systems Design	SBCCI	33
Brazilian Symposium on Software Engineering	SBES	35
Brazilian Symposium on Computer Games and Digital Entertainment	SBGAMES	20
Brazilian Symposium on Programming Languages	SBLP	25
Brazilian Symposium on Formal Methods	SBMF	24
Brazilian Symposium on Software Quality	SBQS	20
Brazilian Symposium on Computer Networks and Distributed Systems	SBRC	39
SIBGRAPI Conference on Graphics, Patterns and Images	SIBGRAPI	33
Symposium on Virtual and Augmented Reality	SVR	23
Brazilian Symposium on Multimedia and the Web	WebMedia	22

*As of March 2022.

Next, we filtered the publications to exclude those with the type listed as "Editorship", to ensure that non-scientific works would not be included in the final analysis. After the filtering process, the data frame contained 6569 publications with 23573 authorships. Table 2 presents information regarding the number of articles and authorships for each of the analyzed conferences.

Table 2. Data collected per conference.

Conference	Total Publications	Filtered Publications		Number of Authorships	Years Indexed on DBLP
		Count	%		
IHC	620	518	83.5	1785	2006, 2008, 2010-2021
SBAC-PAD	711	711	100	2760	2002-2021
SBBD	680	109	16	420	1999-2021
SBCCI	828	827	99.8	2944	2003-2020
SBES	489	489	100	1940	2009-2021
SBGAMES	241	241	100	881	2009-2011, 2014-2015, 2017-2021
SBLP	112	112	100	352	2012-2021
SBMF	178	178	100	500	2009-2018, 2020-2021
SBQS	155	155	100	604	2018-2021
SBRC	603	525	87	1930	2014, 2015, 2017-2021
SIBGRAPI	1231	1231	100	4083	1999-2021
SVR	620	603	97.2	2322	2012-2021
WEBMEDIA	890	870	97.7	3052	2005, 2006, 2008, 2009, 2012-2021

3.2. Gender Inference

Considering that gender information is not available in most databases, including DBLP, one of the most reliable ways to infer an author's gender is by analyzing their name. We combined three different ways to assign gender to first names, the Gender API[6], the Python package gender-guesser[7], and the gender classification data made available

[6]https://gender-api.com/

[7]https://pypi.org/project/gender-guesser/

by the Brazil.IO project[8]. Gender API is an online service with an extensive database. Gender-guesser is an offline package with a more limited amount of names in its dictionary, however, its data was manually checked by native speakers of different countries, and therefore is presumed to be of high quality, as supported by Santamaría and Mihaljevic [Santamaría and Mihaljević 2018].

Brasil.IO's gender classification data reflects the self-informed gender of Brazilian residents as collected in the 2010 Brazilian Census[9], and is made available in a csv file containing the name's classification (male or female), its frequency of appearance as female and male, and the ratio (ranging from 0 to 1), which represents the confidence for the classification. We filtered the data to only include classifications for names that had a ratio of at least 0.9.

We extracted 23573 author names from the list of publications, split them into firs and last names, and used the first name string to first query gender-guesser. Gender-guesser assigns gender as unknown (for a name not found in the database), andy (androgynous names, i.e. names that have a similar probability to be male than to be female), male, female, mostly_male, or mostly_female. Then, for the names that were assigned gender as unknown by gender-guesser, we applied Gender API. Gender API returns a gender assignment with the possible values of male, female or unknown. Lastly, we cross-referenced the names that were still classified as unknown with Brasil.IO's data, applying their classification for names that had a ratio of at least 0.9, in order to avoid ambiguity.

We were able to assign a gender for 91,88% of names in the authorship list. It is important to note that research articles often have more than one author, and albeit co-authors might have different extents of contribution, for the purpose of this study, we considered all authors listed in the publication equally. In order to analyze gender representation in scientific productions, every author name listed was counted as one authorship, meaning that one author could have published more than once in the conferences and, for each instance, it would have counted as different authorships. The number of unique author entries and authorships is shown in Table 3.

Table 3. Unique author entries and authorships by gender assignment.

Gender Assignment	Unique Author Entries	Number of Authorships
Male	8691	17140
Female	2000	4120
Unknown	1271	1914
Mostly_male	189	382
Mostly_female	13	17

By comparing the number of unique author entries with the authorships count, as detailed in Table 3, we define the productivity factors p_f and p_m. p_f, shown in Equation 1, denotes the productivity factor for female authors. p_m, show in Equation 2, denotes the productivity factor for male authors.

[8] https://brasil.io/dataset/genero-nomes/nomes/

[9] https://censo2010.ibge.gov.br/nomes/#/search

$$p_f = \frac{Authorship(f + mf)}{Unique(f + mf)}, \tag{1}$$

where f stands for gender assignment female and mf stands for mostly_female.

$$p_m = \frac{Authorship(m + mm)}{Unique(m + mm)}, \tag{2}$$

where m stands for gender assignment male and mm stands for mostly_male.

3.3. Implementation Details

The main goal for the program's implementation was to make it as accessible and easy to replicate as possible. Our methodology was implemented using Python 3.10[10] on a 11th Gen Intel(R) Core(TM) i5-1135G7 with a 16GB memory. We used the Python Requests library[11] to retrieve conference data from the DBLP API. This information was stored in a csv file containing the associated metadata for each publication. The list was then filtered to remove publications that did not contain DOI information or whose type was listed as "Editorship", as explained in Section 3.1. The data frame used for this analysis is available at GitHub repository[12].

We used the gender-guesser package version 0.4.0, and the Gender API service, following the "Simple Usage" Request, as detailed on the API's documentation [13]. In order to reduce costs with the paid subscription of gender API, every name query and correspondent gender assignment were stored in a JSON file that the program would check before querying to the API, avoiding that multiple requests be made for names who appeared in the list repeatedly.

4. Results

This section presents the results obtained with our proposed methodology. All results are reflective of the DBLP repository as of March 19, 2022. We analyzed conferences ranging from 1999 to 2021, with a total of 6569 publications and 23573 authorships, as shown in Table 2. Although gender inference based on name is not an infallible method, and raises ethical concerns for its exclusion of non-binary individuals, its application was relevant for this study, as we were able to gather gender data that would otherwise be inaccessible.

As the results of this work suggest, despite the number of female authors growing in the past 22 years, as of today, women are still underrepresented in Computer Science research published in SBC conferences, as shown in Figure 1. Furthermore, the gender gap for women in the analyzed conferences is significantly present in all years, as illustrated in Figure 2. Concerning the overall gender representation by conference, as shown in Table 4 and Figure 3, we find that only two conferences, the Brazilian Symposium on Human Factors in Computational Systems (IHC) and the Brazilian Symposium on Software Quality (SBQS), have had at least 30% female authorships in their publications.

[10] https://www.python.org/

[11] https://docs.python-requests.org/

[12] https://github.com/Virtual-Humans-Lab/Gender-Representation-Analysis

[13] https://gender-api.com/en/api-docs#simple-usage

We also found that two conferences, the Brazilian Symposium on Programming Languages (SBLP) and the Brazilian Symposium on Integrated Circuits and Systems Design (SBCCI), have had under 10% female representation.

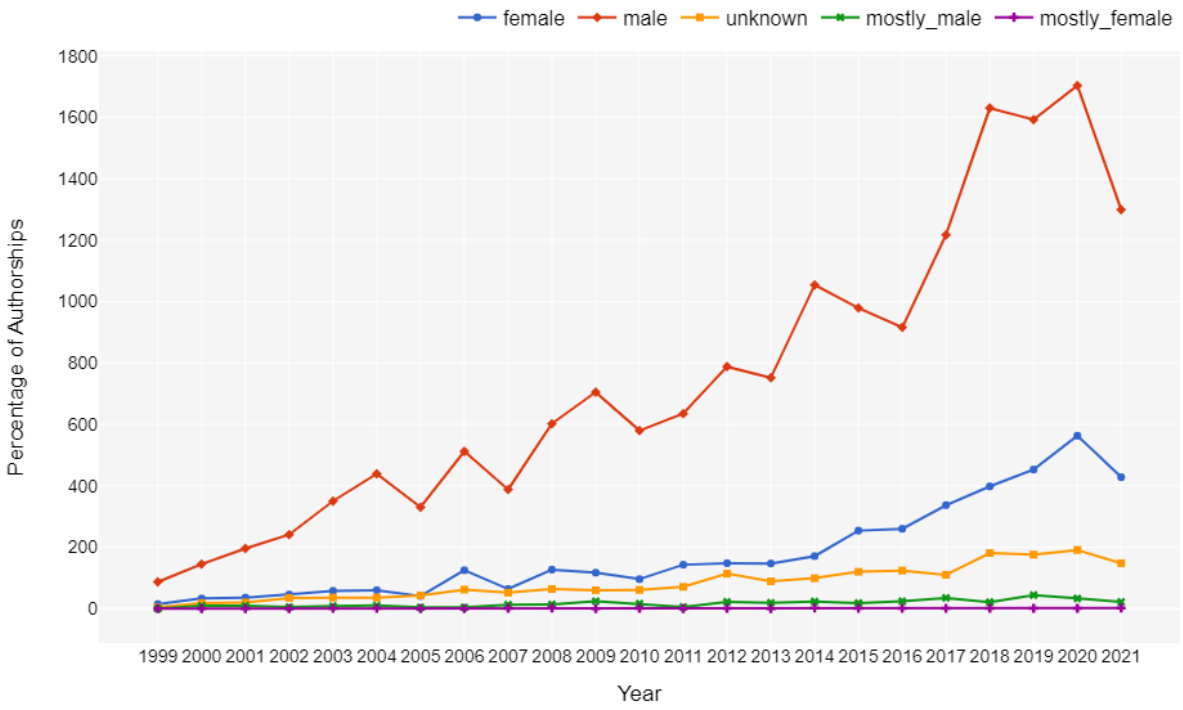

Figure 1. Number of publications by gender per year in SBC's conferences.

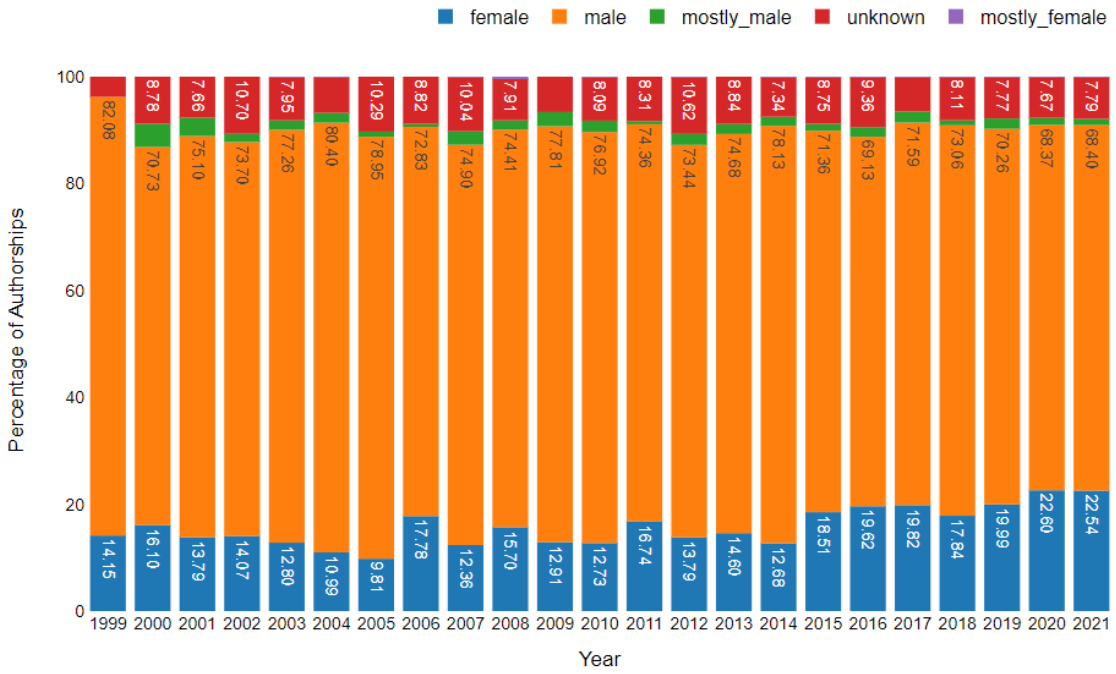

Figure 2. Gender representation by year in SBC's conferences.

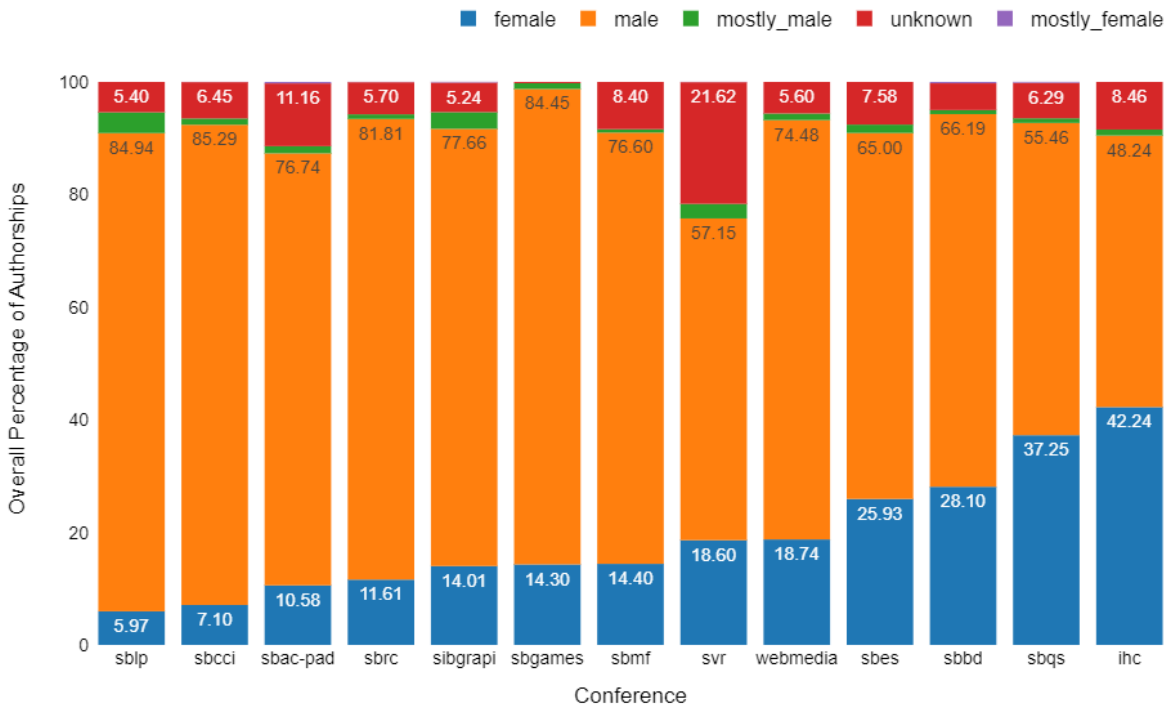

Figure 3. Overall gender representation by conference.

Table 4. Overall gender representation by conference.

Conference	Male		Female		Unknown		Mostly_male		Mostly_female	
	Count	%	Count	%	Count	%	Count	%	Count	%
IHC	861	48.24	754	42.24	151	8.46	19	1.06	0	0
SBAC-PAD	2118	76.74	292	10.58	308	11.16	35	1.27	7	0.25
SBBD	278	66.19	118	28.10	20	4.76	3	0.71	1	0.24
SBCCI	2511	85.29	209	7.10	190	6.45	33	1.12	1	0.03
SBES	1261	65.00	503	25.93	147	7.58	29	1.49	0	0
SBGAMES	744	84.45	126	14.30	2	0.23	9	1.02	0	0
SBLP	299	84.94	21	5.97	19	5.40	13	3.69	0	0
SBMF	383	76.60	72	14.40	42	8.40	3	0.60	0	0
SBQS	335	55.46	225	37.25	38	6.29	5	0.83	1	0.17
SBRC	1579	81.81	224	11.61	110	5.70	16	0.83	1	0.05
SIBGRAPI	3171	77.66	572	14.01	214	5.24	121	2.96	5	0.12
SVR	1327	57.15	432	18.60	502	21.62	60	2.58	1	0.04
WEBMEDIA	2273	74.48	572	18.74	171	5.60	36	1.18	0	0

By analyzing the productivity factors p_f and p_m, shown in Equations 1 and 2, we see that each female author has published an average of 2.06 times in the analyzed conferences, while male authors published an average of 1.97 time. This suggests that the productivity of female authors is on par with that of male authors for the analyzed conferences. The overall gender classifications are show in 3. These numbers indicate that there are 4.16 times more male authors than females, further showing that there is an

urgent need to find strategies and policies to remedy this gender gap.

5. Final Considerations

We performed an analysis of the Computer Science literature output from Brazilian conferences to investigate gender representation. Our results suggest that, although the number of publications authored by women has increased in the past two decades, women are still severely underrepresented in CS research. Our results also indicate that the main issue is not a smaller output of works by female authors, but that there are few women participating in CS research.

To help improve this scenario, different initiatives were created in Brazil to encourage women to join CS communities. The Brazilian Computer Society, for example, sponsors the Meninas Digitais[14] program, which aims to promote technology to girls in elementary school and high school and encourage them to pursue a career in IT.

Some limitations of this research were related to the small sample of data from the analyzed conferences that were indexed in popular CS databases. The Brazilian Computer Society makes available most, if not all, of their conference publications on SBCOpenLib, but unfortunately SBCOpenLib does not have an integrated API to facilitate this type of systematic research.

Another challenge presented was the use of gender inference services, which erase other gender identities by forcing a binary parameter of male or female. To advance more inclusive and accurate research on diversity and representation of minorities, one alternative would be to collect demographic information such as gender by asking for the author's self-identification and including it with the publication's metadata.

As future work, we intend to expand this research with a larger data set and conduct further statistical analysis, as well as compare the results for Brazilian conferences with other international conferences. Most importantly, we aim to use this research as a reference for future analysis of the CS field, evaluating how the current initiatives and efforts being made to encourage women to join CS might impact the gender balance in the future.

The discussion about female representation goes beyond the ethical responsibility of ensuring more equitable gender representation in the future. Diversity is key to advance CS research, as different backgrounds and life experiences present an advantage by giving individuals unique insights and approaches to problem-solving. We hope that this study prompts other works with a focus on analyzing gender disparities in Brazil, and encourages reflection by the community members about the cause of such problems and possible strategies to increase diversity in the Computer Science field moving forwards.

Referências

Albertini, M. K., Backes, A. R., and Sá., A. L. D. (2019). A study of publication trajectories of the brazilian computer science community. *Anais da Academia Brasileira de Ciências*, 91(3).

Arruda, D., Bezerra, F., Neris, V., Toro, P., and Wainer, J. (2009). Brazilian computer science research: Gender and regional distributions. *Scientometrics*, 79:651–665.

[14] `https://meninas.sbc.org.br/`

Cavacini, A. (2014). What is the best database for computer science journal articles? *Scientometrics*, 102:2059–2071.

Cheong, M., Leins, K., and Coghlan, S. (2021). Computer science communities: Who is speaking, and who is listening to the women? using an ethics of care to promote diverse voices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 106–115, New York, NY, USA. Association for Computing Machinery.

Haines, C. D., Rose, E. M., Odom, K. J., and Omland, K. E. (2020). The role of diversity in science: a case study of women advancing female birdsong research. *Animal Behaviour*, 168:19–24.

Holman, L., Stuart-Fox, D. M., and Hauser, C. E. (2018). The gender gap in science: How long until women are equally represented? *PLoS Biology*, 16.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., and Sugimoto, C. (2013). Bibliometrics: Global gender disparities in science. *Nature*, 504:211–3.

Ribeiro, K., Azevedo, J., Maciel, C., and Bim, S. (2019). Uma análise de gênero a partir de dados da sociedade brasileira de computação. In *Anais do XIII Women in Information Technology*, pages 159–163, Porto Alegre, RS, Brasil. SBC.

Santamaría, L. and Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156.

Wang, L. L., Stanovsky, G., Weihs, L., and Etzioni, O. (2021). Gender trends in computer science authorship. *Communications of the ACM*, 64:78 – 84.