Minority Ethnic Groups in Computer Science Research

What is the Bibliography Data Telling Us?

By Swati Agarwal (IIIT Delhi, India), Nitish Mittal (Wadi.com, UAE), Ashish Sureka (ABB Corporate Research Center, India)

Ensuring racial and ethnic diversity in any field including computer science research is important to maintain a positive and prosperous society. Policy makers and government at the national and international level can take appropriate actions and intervene to elevate minorities based on empirical evidence and facts. We present empirical evidence on scholarly paper contributions in the field of computer science by thirteen ethnic groups. We analyze authorship of thousands of papers from a well-known bibliographic database in computer science and various subfields of computer science such as data engineering, software engineering and theory. We present the extent of contribution of various ethnic groups for a period of 17 years and identify ethnic groups having low contribution. We provide answers to several research questions such as author ethnicity distribution across computer science papers, upward and downward trends across years, extent of gender imbalance and the percentage of most prolific authors across various ethnic groups.

Keywords: Bibliometric Analysis, Minority in Computer Science Research, Computer Science and Society, Data Science Ethnicity Analysis, Gender Imbalance
Categories: *K.4 COMPUTERS AND SOCIETY, K.4.2 Social Issues*
Corresponding Author: *Swati Agarwal*
Email: swatia@iiitd.ac.in

Introduction

Racial and ethnic diversity in science and research is important to ensure a broad scientific community and avoidance of lock-in, bias and group thinking[1]. Research shows that several educational institutions support affirmative action claiming that a diverse student body is more educationally effective than a more homogenous one[2]. Similarly, research studies prove that ethnic diversity in a company's corporate board of directors is positively associated with the financial indicators of a firm's performance[3]. There are also several research studies which show there exists disparity and underrepresented minorities in science and engineering disciplines. Hurtado et al. claim that the rates of science

[1] Stirling, Andy. "A general framework for analyzing diversity in science, technology and society." *Journal of the Royal Society Interface* 4, no. 15 (2007): 707-719

[2] Terenzini, Patrick T., Alberto F. Cabrera, Carol L. Colbeck, Stefani A. Bjorklund, and John M. Parente. "Racial and ethnic diversity in the classroom: Does it promote student learning?." *The Journal of Higher Education* 72, no. 5 (2001): 509-531

[3] Erhardt, Niclas L., James D. Werbel, and Charles B. Shrader. "Board of director diversity and firm financial performance." *Corporate governance: An international review* 11, no. 2 (2003): 102-111

baccalaureate completion for underrepresented minority (URM) undergraduates are dismal. They also claim that according to the Center for Institutional Data Exchange Analysis 1999–2000 SMET retention report only 24% of African American, Latino, and Native American students complete a science bachelor's degree in six years in comparison to 40% of White students[4].

Villarejo et al. states that despite significant efforts over the past 30 years by federal government agencies and private organizations, there continues to be a significant underrepresentation of minority scientists engaged in biomedical and behavioral research in the United States[5]. A study by Watson et al. shows that there is a positive influence of ethnic diversity on leadership, group process and performance in learning teams [6]. Arismendi et al. further states that a diverse workforce in science plays a very important role in creating a competitive advantage, problem solving and innovation[7]. They examine the status of gender and race or ethnicity among the US fisheries science workforce and their findings reveal that women and minorities are still a small portion of tenure-track faculty and federal-government professionals. Agarwal et al. perform an exploratory data analysis on the bibliographical dataset of papers in computer science and provide evidence of low participation of women and gender imbalance[8].

Conducting empirical studies on racial and ethnic inequality and diversity in research and scholarly output in the field of computer science is important to understand the extent of underrepresentation of minority scientists in the field. Government agencies can come-up with intervention programs for increasing the representation of minorities and address the disparity problems once they have enough evidence and empirical data on the racial and ethnic imbalance. Based on our literature survey and previous research studies, we observe that there is no empirical study on racial and ethnic diversity in computer science research computed based on authorship in scientific articles. The work presented in this paper is motivated by the need to create awareness about racial and ethnic diversity in computer science research (measured using bibliometric analysis) for the computer science research community as well as policy makers, sociologists and political scientists. We frame several open research questions and analyze data from DBLP (on-line reference for open bibliographic information on computer science journals, conferences and workshops) to answer the stated research questions.

[4] Hurtado, Sylvia, Nolan L. Cabrera, Monica H. Lin, Lucy Arellano, and Lorelle L. Espinosa. "Diversifying science: Underrepresented student experiences in structured research programs." *Research in Higher Education* 50, no. 2 (2009): 189-214

[5] Villarejo, Merna, Amy EL Barlow, Deborah Kogan, Brian D. Veazey, and Jennifer K. Sweeney. "Encouraging minority undergraduates to choose science careers: career paths survey results." *CBE-Life Sciences Education* 7, no. 4 (2008): 394-409

[6] Watson, Warren E., Lynn Johnson, and George D. Zgourides. "The influence of ethnic diversity on leadership, group process, and performance: An examination of learning teams." *International Journal of Intercultural Relations* 26, no. 1 (2002): 1-16

[7] Arismendi, Ivan, and Brooke E. Penaluna. "Examining diversity inequities in fisheries science: a call to action." BioScience (2016)

[8] Agarwal, Swati, Nitish Mittal, Rohan Katyal, Ashish Sureka, and Denzil Correa. "Women in computer science research: what is the bibliography data telling us?." ACM SIGCAS Computers and Society 46, no. 1 (2016): 7-19

Experimental Dataset

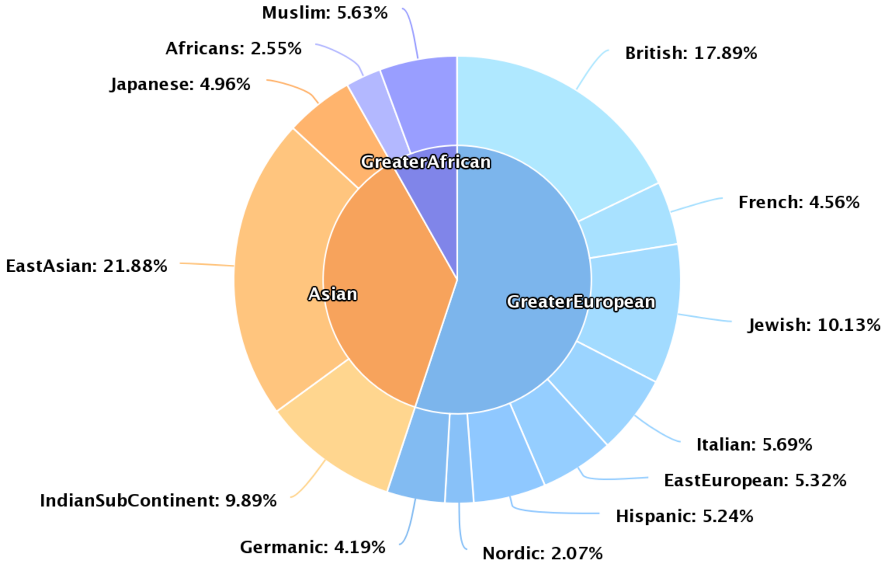

Figure 1 Ethnicity distribution for the articles in the experimental dataset

Our experimental dataset consists of entries extracted from DBLP. The total number of articles that we analyze in the CS, DE, SE and TH domain is 82,603, 19896, 12909, and 23821 respectively. The list of conferences in the DE domain are 17. The list of conferences in SE are 11. The list of conferences in TH are 27. The total number of unique authors in the dataset is 116,850. The average number of authors per paper in the dataset is 3.14. Amongst the authors for which we were able to calculate the gender, the percentage of male and female authors are 77.8% and 22.2% respectively. Figure 1 shows the ethnicity distribution for authors across all the articles in our experimental dataset.

Ethnicity and Gender Classifier

We use two tools for computing the ethnicity and gender of an author. The Name Ethnicity Classifier is a tool developed by Ambekar et al. and publicly available at http://www.textmap.com/ethnicity/[9]. The Name Ethnicity Classifier aggregate entities into different ethnic groups and our analysis is limited by the ethnic groups defined and computed by the tool. The classifier by Ambekar et al. was trained on data obtained from Wikipedia and was available via Web API. We used the Web API of the tool and made a POST request passing the name of the author for which the ethnicity needs to be determined. The tool returned a JSON response consisting of the hierarchical structure of the ethnicity. We extracted the leaves of the hierarchical structure which consists of ethnicity such as British, Jewish, Africans and Muslims. We used the Genderize.io API available at https://genderize.io for determining the gender of an author. As of 16 April

[9] Ambekar, A., Ward, C., Mohammed, J., Male, S., and Skiena, S. 2009. Name-ethnicity classification from open sources. In Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Paris, France, June 28 - July 01, 2009). KDD '09. ACM, New York, NY, 49-58

2017, the Genderize.io database contains 216286 distinct names across 79 countries and 89 languages. We used the Genderize.io Web API and made a GET request passing the name of the author for which the gender needed to be determined. The tool returns a JSON response mentioning the gender of the name.

Empirical Analysis and Results

We frame five Research Questions (RQ) and provide answers in this section.

RQ1: What is the Authors Ethnicity Distribution in Computer Science Research Papers?

We compute the number of papers published by each ethnic group in Computer Science (CS) and three domains within Computer Science: Data Engineering (DE), Software Engineering (SE) and Theory (TH). Figure 1 displays two bar charts. The top bar chart in Figure 2 is for CS whereas the bottom bar chart is for DE, SE and TH. Figure 2 reveals the distribution of the number of articles in the dataset across 13 ethnic groups. Figure 2 reveals that the ethnic group Nordic and African have the least number of articles published in CS, DE, SE and TH. We conclude from Figure 2 that Nordic and African are the minority ethnic group in terms of research paper publications with respect to the 13 ethnic groups considered in our study. We observe that the highest number of articles published in TH are by Jewish but the highest number of articles published in SE are by British followed by EastAsian and then Italian.

The highest number of articles published in DE is by the ethnic group EastAsian. The number of articles in CS published by the 13 ethnic groups in decreasing order are: EastAsian, British, IndianSubContinent, Jewish, Italian, Muslim, EastEuropean, Japanese, Hispanic, Germanic, French, African and Nordic. Our analysis reveals that the number of articles published by EastAsian in CS is 2.5 times the number of articles published by the ethnic group IndianSubContinent. Similarly, the number of articles published by Jewish in CS is 2.2 times the number of articles published by the ethnic group Muslim. We observe that in CS the number of articles published by Japanese, Hispanic, Germanic and French are nearly the same (within 1%-2% of each-other). In CS, the ranking of French is 11th whereas in TH the rank for French is 6th. The rank of EastEuropean is 4th is TH but relatively lower in SE (ranked as 9th).

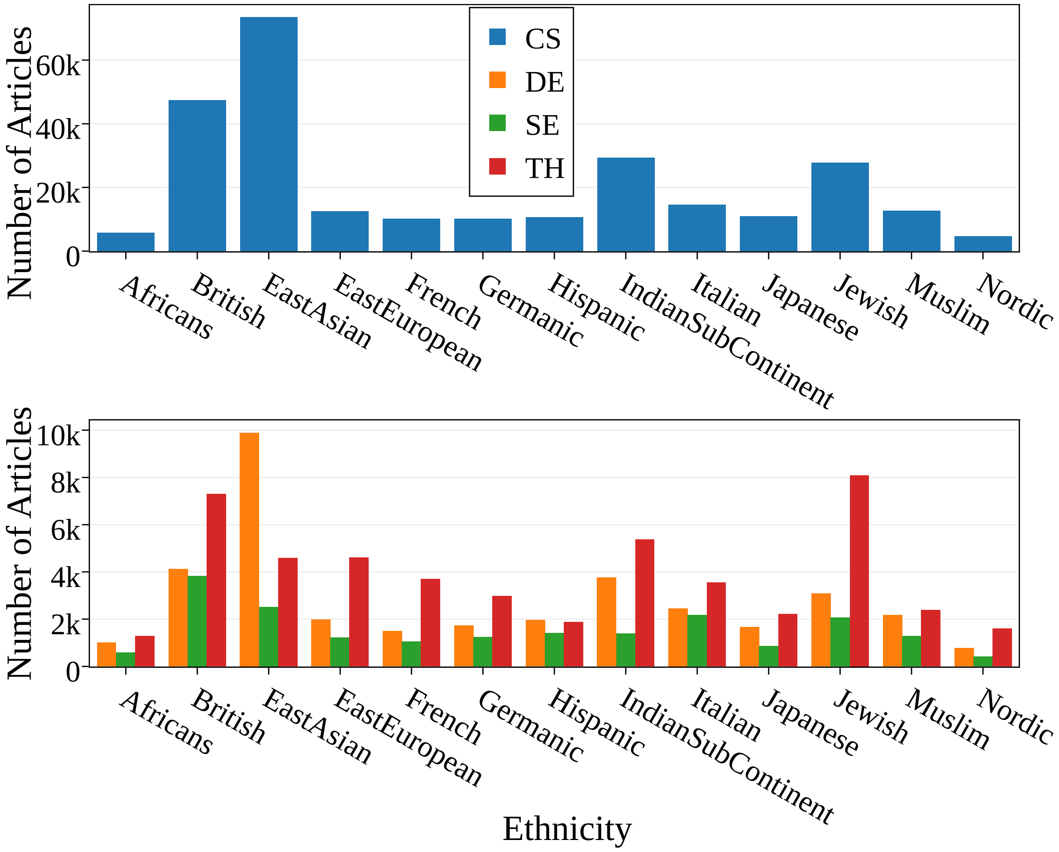

Figure 2 Distribution of Authors' Ethnicity Publishing Articles in Computer Science Research (CSR) Including Data Engineering, Software Engineering and Theory Domains.

RQ2: Is there an Upward or Downward Trend for Minorities and other Ethnic Groups across Years?

In RQ1, we analyzed the distribution of articles across the 13 ethnic groups and in this RQ our objective is to investigate trends in the number of articles across years for the various ethnic groups. Figure 3 and 4 shows a heat-map in which the color of the cells is a reflection of the number of articles published in CS by the ethnic group and year represented by the cell. We draw two heat-maps to bring more clarity and differentiation between the cells in the second heap-map as there is a wide variance in the number of articles published across the 13 ethnic groups. We conclude from RQ1 that the ethnic group Nordic and African are the minority with respect to the 13 ethnic groups considered in our study. Figure 4 shows that for both the ethnic groups (Nordic and African), the number of articles have increased from the year 2000 to 2016. The year 2008 is a mid-point between the year 2000 and 2016. Our analysis reveals that the number of articles published in CS by the ethnic group African in 2008 is 2.6 times than the number of articles published by African in 2000. The number of articles published by African in 2016 is 1.6 times the number of articles published by African in 2008.

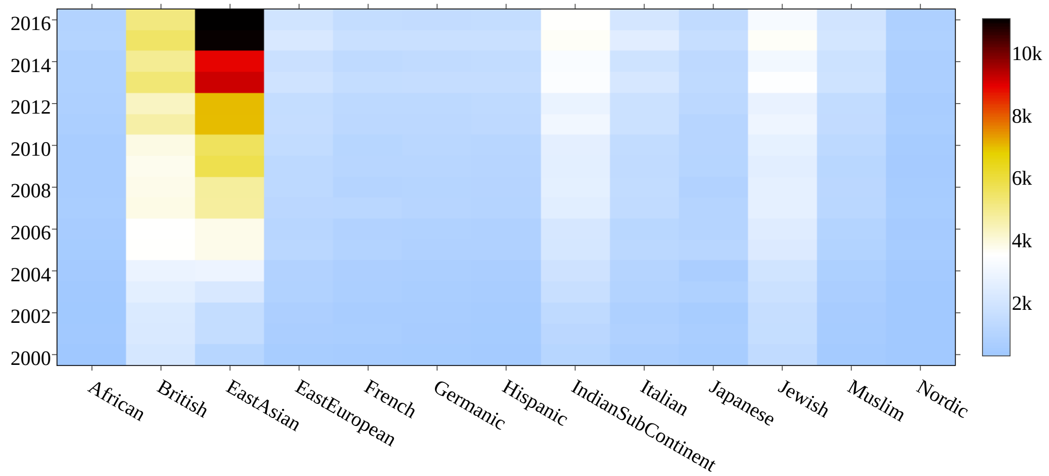

Figure 3 A Timeline Based Review of Variation in Number of Articles Published by Different Ethnic Groups and Minorities

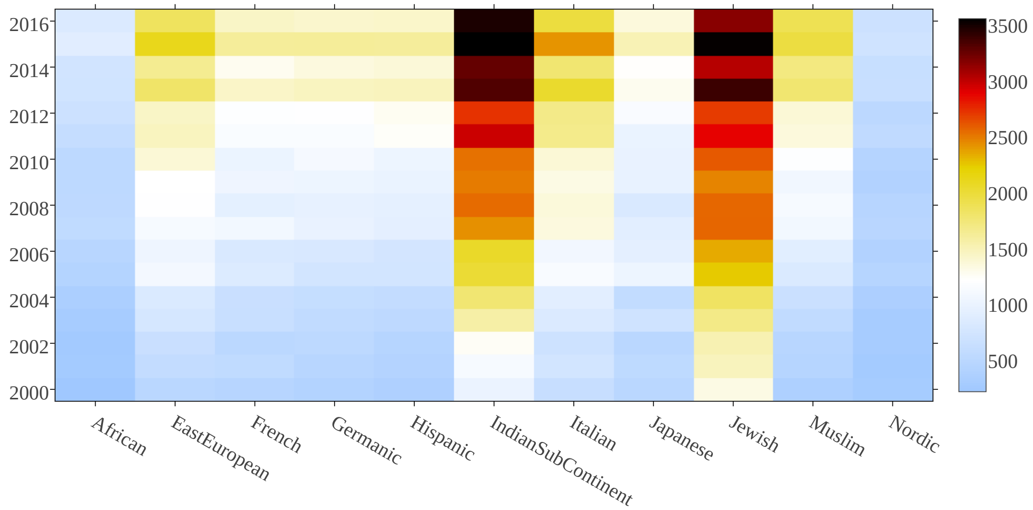

Figure 4 A Timeline Based Review of Number of Publication by Various Minorities and Ethnic Groups Excluding British and East Asian

Our analysis reveals that the number of articles published by Nordic in year 2008 is 1.7 times by the same ethnic group in 2000. The number of articles published by Nordic in year 2016 is 1.5 times the number of articles published by Nordic in year 2008. Amongst the two minorities African and Nordic, the growth in African (3.6 times) is relatively much higher than Nordic (2.4 times). We observe that the ethnic group EastAsian shows the highest growth (7.7 times) from the year 2016 to the year 2000 followed by Muslim for which the growth is 4.1 times for the same period. The lowest growth is observed for Jewish which 2.1 times from the year 2000 to 2016. Our analysis shows that every ethnic group has shown growth in terms of the number of articles published in CS from 2000 to 2016 but the extent of growth varies. The ranking in terms of growth from 2000 to 2016 in decreasing order is: EastAsian (7.7 times), Muslim (4.1 times), African (3.6 times), Hispanic (3.2 times), EastEuropean (2.9 times), French (2.9 times), Italian (2.9 times), Japanese (2.8 times), IndianSubContinent (2.8 times), Germanic (2.75 times), Nordic (2.4 times), British (2.3 times) and Jewish (2.1 times).

RQ3: Is there a Gender Imbalance in Research Papers by Minorities and other Ethnic Groups?

Figure 5 shows a stacked bar chart displaying the percentage of male and female authors across various domains. Our analysis reveals that the percentage of female authors in the Africans ethnic group varies from a minimum of 8.30% to a maximum of 21.30% (from the year 2000 to 2016). As shown in Figure 5, the representation of women authors is highest in the EastAsian. In the EastAsian ethnic group, the percentage of female authors ranges from a minimum of 29.89% to a maximum of 36.77%. We observe that the representation of women authors in Germanic is even lower than that of Africans. In Germanic, the percentage of women authors varies from a minimum of 7.78% to a maximum of 13.87%. In Nordic, the percentage of women authors vary from a minimum of 8.71% to a maximum of 15.28%. The range for the Muslim ethnic group is between 12.62% and 20.79%.

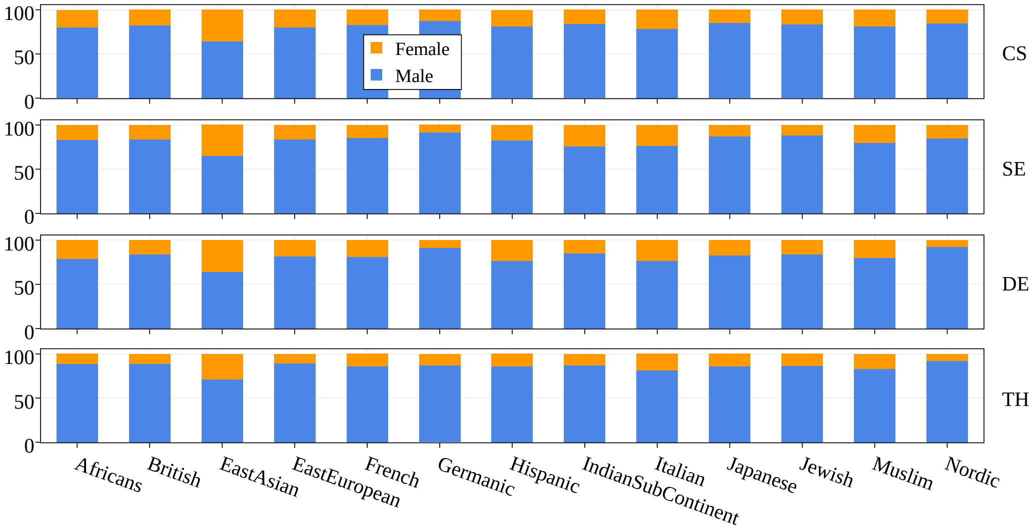

Figure 5 Distribution of Male and Female Authors Publishing Research Papers in CSR Conferences from Minority and Other Ethnic Groups

We conduct an analysis of the percentage of male and female authors for all the four domains in our dataset (CS, DE, SE, TH). We observe that in the CS domain, the percentage of female authors is highest in EastAsian (35.78%) followed by Italian (21.76%) and EastEuropean (20.23%). In the CS domain, the ethnic groups with least percentage of women authors are Germanic (12.57%), Japanese (15.14%) and Nordic (15.43%). In the DE domain also, EastAsian (36%) and Italian (23.4%) are in the top 2 followed by Hispanic (23.22%). In DE, Nordic (7.5%) and Germanic (8.57%) are the only two ethnic groups with percentage of female authors less than 10%. In SE domain, the percentage of female authors varies from a minimum of 8.47% for Germanic and a maximum of 35.07% for EastAsian. In SE domain, the ethnic groups with percentage of female authors above 20% are: EastAsian (35.07%), IndianSubContinent (23.95%), Italian

(23.44%) and Muslim (20.28%). In TH domain, Nordic (7.79%) is the only ethnic group with percentage of female authors less than 10%. Our analysis reveals that in TH domain, only three ethnic groups have percentage of female authors above 15%: EastAsian (28.7%), Italian (18.75%) and Muslim (16.8%).

RQ4: Is there an Upward or Downward Trend on Gender Imbalance in Research Papers by Minorities and other Ethnic Groups?

Figure 6 and 7 shows a heat-map for all the ethnic groups in our dataset for the period from 2000 to 2016. We observe that there is an increasing trend in the percentage of female authors for the Africans ethnic group. The percentage of female authors is below 19.0% before 2009 but from 2000 onwards the percentage of female authors is always above 19.0%. The highest percentage of female authors is 21.30% in 2016 which is the recent most year with respect to the study presented in this paper. As shown in Figure 6 and 7, we do not observe any upward or downward trend across the 17 year period for British. The percentage of female authors for British in year 2001 was 18.46% and it is approximately the same in the year 2016 also (18.95%). We observe a slight increase in the percentage of female authors for the EastAsian ethnic group. The percentage of female authors for the EastAsian is within the 29% to 32% range in the 2000 to 2003 period whereas it is in the 34% to 36% range within the 2007 to 2016 period.

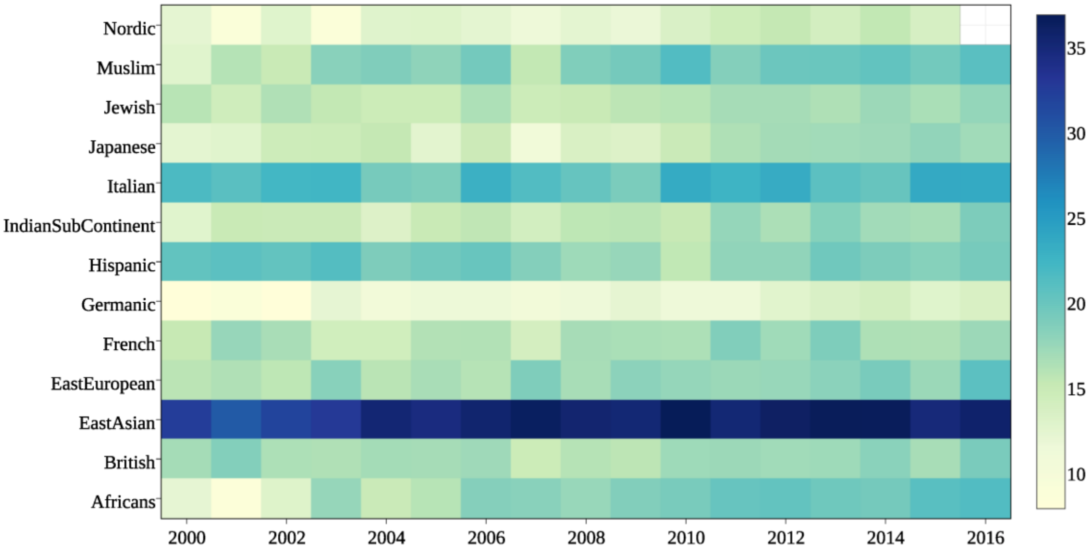

Figure 6 Timeline Based Review of Female Authors in Minority and Other Ethnic Groups

The EastEuropean ethnic group shows an overall increasing trend with a small number of declines in between for few years. The percentage of female authors is in the 15% to 16% range for the year 2000 – 2002. However, since 2010 the percentage of female authors is always above 17%. The percentage of female author for the year 2016 is 20.65%. In the case of French ethnic group, we do not see any overall increasing or decreasing trend, the percentage of female authors is around 16% to 17% throughout the 17 year period except a small number of minor deviations. The Germanic ethnic group shows a clear upward trend with the percentage of female authors below 9% from 2000 to 2002 but increasing to more than 12.5% consistently after 2012. Similarly, we observe an overall increasing trend for the IndianSubContinent ethnic group in which the percentage of female authors increases from 12% - 14% in the year 2000 to 2005 to 15% to 18% in the year 2006 to 20016. We do not observe any upward or downward trend for the Hispanic and Nordic ethnic group. We observe an upward trend for the Muslim ethnic group. For the year 2000 and 2002 the percentage of female authors for the Muslim ethnic group is 12.62% and 14.84% respectively but the percentage value is consistently above 19.0% after the year 2012 (i.e., from the year 2012 to 2016).

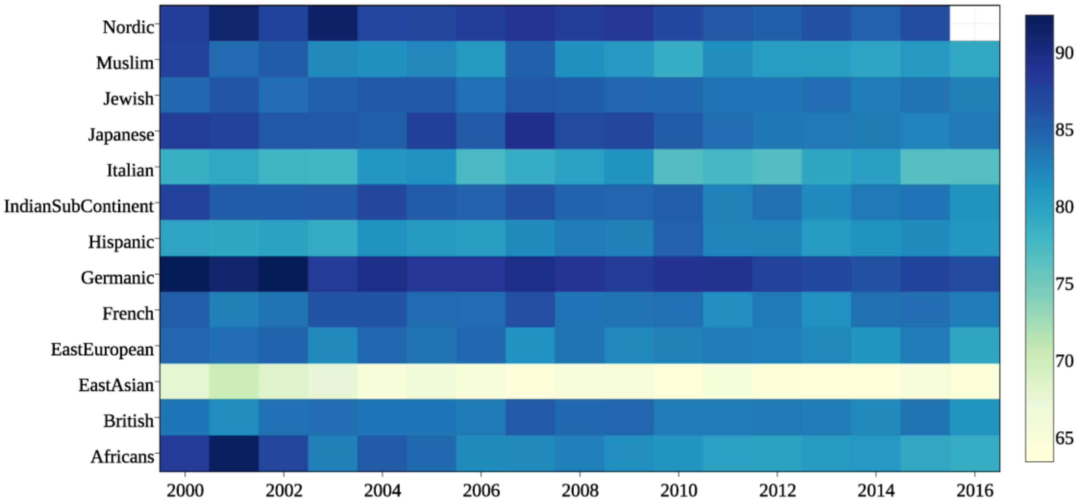

Figure 7 Timeline Based Review of Male Authors Publishing Papers in CSR Conference from Minority and Other Ethnic Groups.

RQ5: What is the percentage of Minorities and other Ethnic Groups in Top 50 most prolific authors within a particular field?

We identify the top 50 most prolific authors (authors who have published maximum number of papers) in our dataset across all the four domains. We compute the ethnicity of the top 50 most prolific authors. Figure 8 shows the bar graph where the x-axis represents the ethnicity and the y-axis represents the frequency or the number of times authors from that ethnicity are in the top 50. Figure 8 shows that 72% of the 50 most prolific authors in CS are from EastAsian and British ethnicity. In CS, there are 3 authors each from French, IndianSubContinent and Jewish ethnicity. There are no authors from EastEuropean, Hispanic and Nordic ethnicity in CS. Our analysis reveals that in DE, 20% of the 50 most prolific authors are from IndianSubContinent which is not the case in CS. The top 3 ethnicities in DE are EastAsian, IndianSubContinent and British. The rank of British in CS is second whereas in DE the rank of British is third. It is interesting to note that while in CS and DE there are no authors from EastEuropean ethnicity, 10% of the 50 most prolific authors in SE and 8% of the 50 most prolific authors in TH are from EastEuropean ethnicity. Figure 8 shows that the trend in SE is different than the trends in CS and DE. In SE, 26% of the 50 most prolific authors are British followed by Italian which is 16%. The Italian ethnicity is more prominent in the list of the 50 most prolific authors in the domain of SE in comparison to CS, DE and TH.

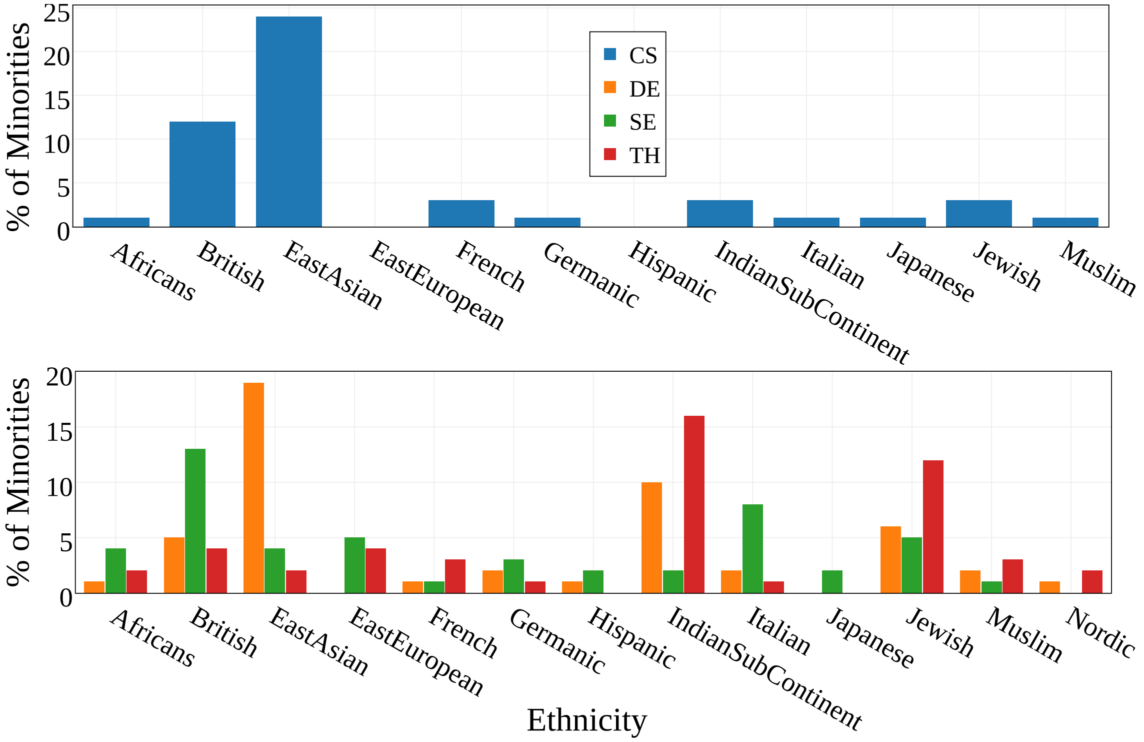

Figure 8 Distribution of Minorities and Other Ethnic Groups in Top 50 Prolific Authors within CSR, Data Engineering, Software Engineering and Theory Domains

Our analysis reveals that French, Muslim and Nordic are minorities in CS as they constitute less than 3% of the 50 most prolific authors. TH domain shows a different trend in comparison to other domains. In TH, the ethnicity which is in the list of 50 most prolific authors is IndianSubContinent (32%) followed by Jewish (24%). The Jewish ethnicity has the highest prominence in the TH domain in context to the most prolific authors. The minorities in the TH domain are Germanic, Hispanic, Italian and Japanese as they constitute less than 3% of the most prolific authors. In general and taking all the four domains into consideration, we notice that the EastAsian, British, IndianSubContinent and Jewish are amongst the most prolific authors. On the other hand, we notice that combining all four domains, Hispanic, Japanese and Nordic are in the minority. Our analysis reveals that the representation of Africans is slightly above Hispanic, Japanese and Nordic across all four domains.

Conclusion

Our analysis reveals that the ethnic group Nordic and African have the least number of articles published in Computer Science (CS), Data Engineering (DE), Software Engineering (SE) and Theory (TH). Nordic and African are the minority ethnic groups in terms of research paper publications with respect to the 13 ethnic groups considered in our study. We observe that the ethnic group EastAsian shows the highest growth (7.7 times) from the year 2000 to the year 2016 followed by Muslim for which the growth is 4.1 times for the same period. The lowest growth is observed for Jewish which is 2.1 times from the year 2000 to 2016. The number of articles published by African in 2016 is 1.6 times the number of articles published by African in 2008. The representation of women authors is highest in EastAsian. In the EastAsian ethnic group, the percentage of female authors ranges from a minimum of 29.89% to a maximum of 36.77%. The representation of women authors in Germanic is lowest in comparison to all the ethnic groups in our dataset. In Germanic, the percentage of women authors varies from a minimum of 7.78% to a maximum of 13.87%. We observe that there is an increasing trend in the percentage of female authors for the Africans, IndianSubContinent, EastAsian and EastEuropean ethnic group. Our results indicate that 72% of the 50 most prolific authors in CS are from EastAsian and British ethnicity. Our analysis reveals that French, Muslim and Nordic are minorities in CS as they constitute less than 3% of the 50 most prolific authors. In SE, 26% of the 50 most prolific authors are British followed by Italian which is 16%.

We conclude that there is significant gender imbalance in-terms of paper published by male and female authors. The percentage of papers published by female authors across all ethnic groups is significantly lower than the percentage of papers published by male authors. Furthermore we do not observe any increasing trend in terms of contributions from female authors except minor ups and downs across years. Similarly, we conclude that there is an imbalance in-terms of the contribution from various ethnic groups and the disparity between the highest and lowest is significant. We observe both kind of phenomenon: some ethnic group showing an increasing trend whereas some ethnic groups either decreasing or remaining more or less constant across several years.