FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Department Informatik

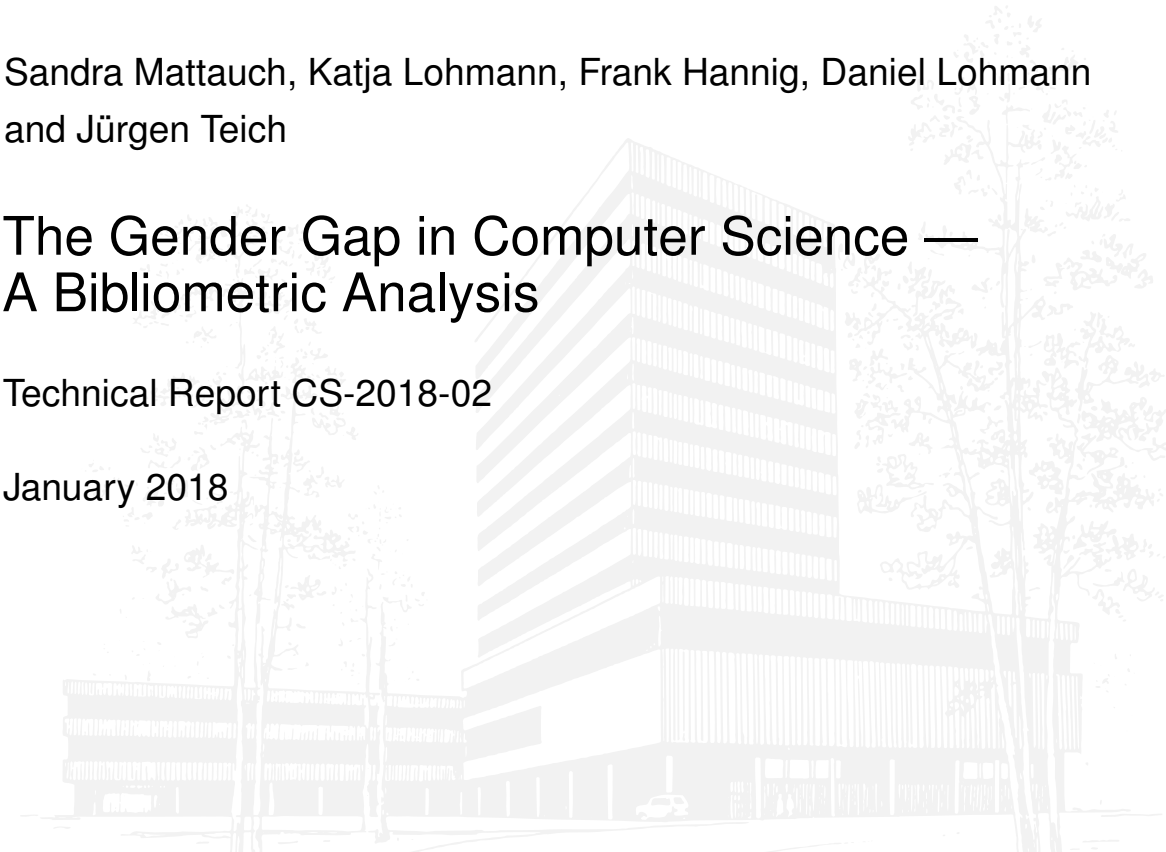
Technical Reports / ISSN 2191-5008

Sandra Mattauch, Katja Lohmann, Frank Hannig, Daniel Lohmann and Jürgen Teich

The Gender Gap in Computer Science — A Bibliometric Analysis

Technical Report CS-2018-02

January 2018



Please cite as:
Sandra Mattauch, Katja Lohmann, Frank Hannig, Daniel Lohmann and Jürgen Teich, “The Gender Gap in Computer Science — A Bibliometric Analysis,” Friedrich-Alexander-Universität Erlangen-Nürnberg, Dept. of Computer Science, Technical Reports, CS-2018-02, January 2018.

Friedrich-Alexander-Universität Erlangen-Nürnberg
Department Informatik

Martensstr. 3 · 91058 Erlangen · Germany

www.cs.fau.de

The Gender Gap in Computer Science — A Bibliometric Analysis

Sandra Mattauch, Katja Lohmann, Frank Hannig, Daniel Lohmann and Jürgen Teich
Computer Networks and Communication Systems
Dept. of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
sandra.mattauch@fau.de

***Abstract*— The low share of women in computer science is documented by many surveys. Most of these studies are based on registrations or enrolments of universities or other scientific institutions. In this paper, we present a new approach to a) analyse the gender gap in the group of scientists that are currently active in research and b) classify differences for different fields of computer science. This group comprises professors, industrial researchers, senior lecturers, postdoctoral researchers, and doctoral students shortly before finishing their theses. The proportion of women in a specific scientific area of computer science might provide valuable information for strategies to recruit women as postdocs or professors.**

I. Introduction

Women are underrepresented in the STEM fields (Science, Technology, Engineering, and Mathematics) — in Germany, but also in other parts of the world. This was demonstrated in several surveys investigating the proportion of women in the STEM fields for specific populations. Some of these studies, for example, investigated the number of enroled students ([6], [14]) or the percentage of female professors at universities. Other studies analysed the disparities in research funding [10]. Nearly all surveys selected a special population of women in consideration of their university degree or their nationality (e.g. [12], [17]). These surveys are usually based on data records from several kinds of registrations or enrolments, for example, the enrolment as student or doctoral student, or the registration of finished doctoral theses. But especially researchers at the postdoctoral level or industrial researchers are often not registered and unfortunately drop out of the surveys. Additionally, these surveys do not provide information about the scientific activity of a researcher.

In this study, we present a method to analyse the gender gap in the group of scientifically active researchers regardless the limitations mentioned above, and focussed on a certain scientific field. The group of interest comprises scientists that are currently active in doing research and publishing their findings — regardless of their university degree, their nationality, gender, age, or origin and irrespective of their employment level in university or industry. As a case study, we investigate the gender gap in the scientific field of the CRC/Transregio 89 *Invasive Computing*[1] which covers research from diverse domains of computer science and electrical engineering such as computer engineering, operating systems, programming languages, security, robotics, and high-performance computing. To ensure that only scientifically active scientists are taken into account, we decided to collect data of researchers that successfully published their results in the proceedings of international conferences within the last three years. Conferences and the appropriate conference proceedings are the common publication medium in computer science and have a much higher impact than journal papers. For this purpose and for working with representative and high quality data, we used the DBLP Computer Science Bibliography [4], which lists the major computer science journals and conference proceedings, as database. Table I presents a summary and selection of 19 most relevant conferences for different disciplines of computer science according to scientists of our transregional research centre. Based on this selection, we developed a Perl script

[1] http://invasic.informatik.uni-erlangen.de

extracting the author names by the given constrains (conference name and a time period of three years). Based on the filtered results, we subsequently determined the gender and country origin of each author by NamSor Applied Onomastics [8]. We finally verified this approach by random sampling and manual classification of the sampled names. The extracted information was then used to analyse the gender gap in the field of the CRC/Transregio 89 *Invasive Computing*.

Table I

List of Conference
ACNS
ARCS
ASAP
ASP-DAC
ASPLOS
CASES
CC
CODES+ISSS
DAC
DATE
Euro-Par
Eurosys
PARCO
SOSP
USENIX
VEE
DASIP
Humanoids
NDSS

To the best of our knowledge, this is the first report that presents a method to analyse the gender gap in an unlimited population (independent of degree, gender or origin) that could be narrowed down to a scientific field of interest in any time period.

II. Methods

A. Extraction of author names from the DBLP Computer Science Bibliography

To gather the original population of all scientifically active researchers within the field of the CRC/Transregio 89, we extracted the names of authors contributing to most relevant conferences (Table I) within the last three years from the DBLP Computer Science Bibliography [4].

The DBLP Computer Science Bibliography provides bibliographic information on all major computer science journals and proceedings. This open-data service indexes more than 3 million articles, published by more than 1.4 million authors [4].

To extract the author names from the DBLP database, we created a Perl script. This script extracts all author names — regardless of the order of authors — for all papers published at a certain conference. The conference is defined by the input variables *venue* and *year*. The *venues* are the acronyms of the conferences as listed in Table I. For *years* we chose the recent years 2012, 2013, and 2014. The script displays a list with the authors' first and last name, and the conference name and year. The resulting population comprises 10,003 authors.

B. Data handling

The extracted author names from the DBLP database were subsequently classified by NamSor Applied Onomastics, a name recognition software provided by a private start-up company [8]. The specialised data mining software also recognises the linguistic or cultural origin of each personal name in any alphabet/language and allocates an onomastic class and the gender to each author name. The innovative machine learning algorithm provides an unmatched accuracy at a fine grain level, flexibility and integration capability, to filter through large databases and extract names. It recognises which language or culture stands behind a given name [8].

To ensure a high degree of accuracy in the classification of the author names, we decided to use NamSor Origin API first, followed by NamSor Gender API.

Determination of the likely country of origin of a name by NamSor Origin API: NamSor Origin API allows to determine the likely country of origin of each author, based on the sociolinguistics of the name (language, culture). The anthroponomical classification can be summarised as follows: Judging from the name only and the publicly available list of all 150k Olympic athletes since 1896 (and other similar lists of names), for which national team would the person most likely run? Here, the United

States, Australia, etc. are typically considered as a melting pot of other cultural origins (Ireland, Germany, etc.) and not as an onomastic class on its own [13].

Table II: Onomastic Classes (top 20)

Onomastic Class	Percentage	Total
China	11.4 %	1140
India	8.8 %	882
Germany	8.3 %	832
France	6.8 %	676
United Kingdom	6,6 %	657
Italy	5.7 %	566
Japan	4.9 %	495
Taiwan	4.8 %	478
Spain	3.5 %	352
Republic of Korea	3.3 %	332
Greece	2.6 %	259
Switzerland	2.5 %	249
Iran	2.2 %	222
Austria	2.0 %	206
Ireland	1.6 %	159
Pakistan	1.5 %	151
Romania	1.4 %	137
Russian Federation	1.2 %	118
Belgium	1.1 %	112
Netherlands	1.1 %	111

Based on the NamSor Origin API algorithm, the basic population of 10,003 authors was classified into 89 onomastic classes. Table II presents the 20 proportionally largest classes, which represent 81.3 percent of the basic population. 26 onomastic classes have less than 10 authors listed and represent together under one percent of the basic population.

Determination of the likely gender of a name by NamSor Gender API: Next, we determined the likely gender of a personal name by using the NamSor Gender API. The software predicts the gender of a personal name on a -1 (male) to +1 (female) scale and covers the US, European, Indian, African, Chinese, Hebrew, Russian/Slavic/Cyrillic, and Arabic names. In this step, the software combines two algorithms to maximise accuracy. First, a unique global name sociolinguistics algorithm, which (1) recognises the origin of the couple first name and last name and (2) infers whether the name sounds male or female in that particular culture. Second, a query in a massive database (800,000 names), which contains statistical information about baby names in each country of the world [7]. Nevertheless, NamSor recommends to pass additional geography/local context to the names to improve the accuracy of classification [7]. The reliability of this method was already investigated in several publications [15], [13], [1] [2].

Figure 1 reveals that 79.8 percent of the author names are classified as male, and only a small proportion of 11.8 percent are classified as female names. 8.4 percent of the names in the basic population are unclassified (scale 0). These not classified names mainly have three reasons: 1) The usage of initials instead of the full first name, 2) names like Kerry, Jean, or Maria that are not strongly correlated to gender, and 3) the structure and usage of Asian names. Because it is not possible to reliably determine the gender for these unclassified names, we excluded this group of 841 names from the following analyses.

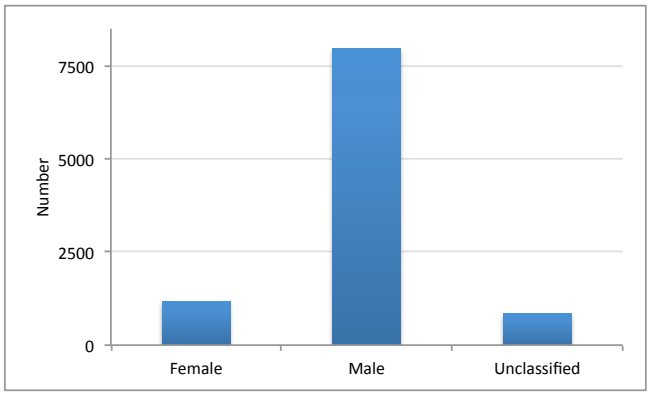


Figure 1: Distribution of female, male, and unclassified names as assorted by NamSor Gender API in the original population.

Removal of Asian names: In most countries and cultures, the method of onomastics is very accurate, with a precision in the range of 95-99 percent — but we should pay attention to the structure of Asian names. The used Perl script generates a list of authors with first name and family name. In Asia, the family name comes first, followed by the first name. Although there are currently over 4,000 Chinese surnames, only hundred surnames still make up over 85 percent of China's 1.3 billion citizens. In fact, just the top three Wang, Li and

Zhang cover more than 20 percent of the population [9]. The situation is aggravated by the fact that a lot of Chinese names are not strongly correlated to gender. And if they were transliterated in Latin characters even more information gets lost. The automatic determination of gender from Asian names with sufficient accuracy is not within the bounds of possibility of this work. Even some experts of onomastic say "we recommend tossing a coin instead" [7]. For these reasons, we decided to exclude these names from the onomastic classes listed in Table II. Removal of all names from these onomastic classes reduces the population by 1,444 to 7,718 names.

Table III: List of Excluded Classes

Onomastic Classes
Hong Kong
China
Taiwan
Republic of Korea
Viet Nam
Democratic People's Republic of Korea

In Figure 2, the distribution of male, female, and unclassified authors after the removal of Asian names is shown. The percentage of female authors remains almost unchanged, but the amount of unclassified names has been reduced to 2.7 percent. The number of male authors has increased accordingly to 86.2 percent.

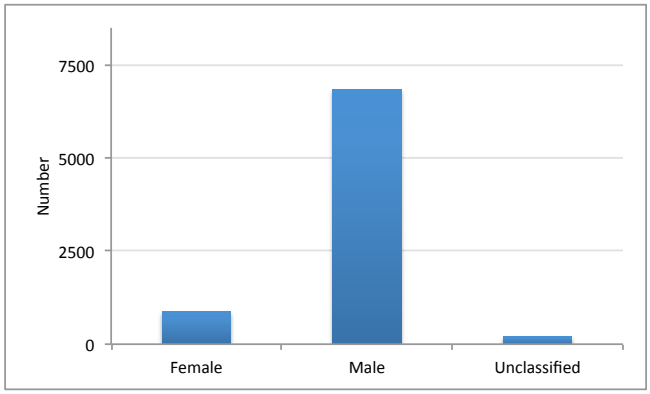


Figure 2: Distribution of female, male, and unclassified names as assorted by NamSor Gender API in the population when disregarding Asian names.

C. Validation of name sorting

After applying the procedure described in Section II-B, we ended up with a population of 7,718 names (basic population): 878 names were classified as female names, 6,840 as male names. To test whether the names classified as female names really belong to women and — vice versa — those classified as male names really belong to men, we randomly selected samples from the basic population of men and women. The minimal sample sizes n of women and men is calculated using the following formula:

$$n \geq \frac{N}{1 + \frac{(N-1)*e^2}{z^2*P*(1-P)}} \quad (1)$$

In Equation (1), N is the number of elements in the stock population, e the margin of error (5%), z is the z-score (1.96 for a confidence level of 95%), and P the prior judgement of the correct distribution (0.5, no prior judgement).

This gives us a sample size of 268 for the group of female names and 364 for the group of male names. The gender of scientists from these sample groups was manually verified by searching them on the internet — assuming scientifically active persons to have an internet presence. We determined the gender of the scientists by photos and the usage of gender-specific keywords (he, she, him, her, etc.) on the personal homepages, on platforms like LinkedIn [5] or ResearchGate [11], or on pages referring to the scientist, for instance, as authors.

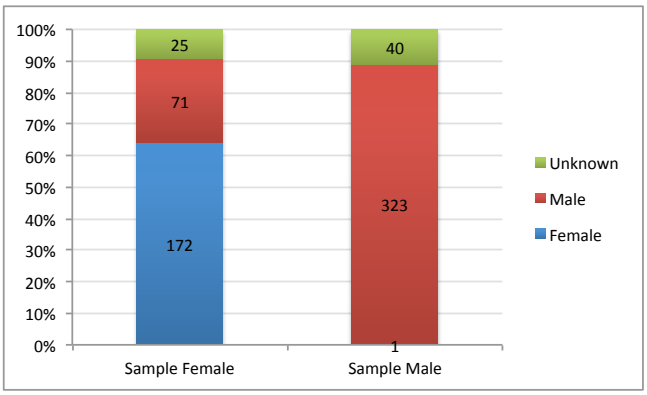


Figure 3: Verification of gender in the sample of female and male names.

The results are shown in Figure 3. The estimation

of the likely gender of a person by "NamSor Gender API" works quite well for male scientist but noticeably not as good for the group of female scientists (see Figure 3): In the group of men, 89 percent were correctly verified to be male, only 0.3 percent were female, and 10 percent could not be verified due to no internet presence. In the group of women, only 64 percent were correctly verified as female, yet 27 percent as male and 9 percent can not be found on the internet.

Based on these random experiments, we decided to correct the automatically determined number of female and male authors accordingly using the following term:

$$F_{corr} = F * corr_{ff} + M * corr_{fm} \quad (2)$$

$$M_{corr} = M * corr_{mm} + F * corr_{mf} \quad (3)$$

In Eqs. (2) and (3), F_{corr} and M_{corr} denote the corrected numbers of women and men, F and M are the original values obtained from the name-sorting procedure (Section II-B), and $corr_x$ are the correction factors estimated from the results of the verification of name sorting:

$$\begin{aligned} corr_{ff} &= \text{females in female group} &&= 0.64 \\ corr_{mm} &= \text{males in male group} &&= 0.89 \\ corr_{fm} &= \text{females in male group} &&= 0.3 \\ corr_{mf} &= \text{males in female group} &&= 0.27 \end{aligned}$$

The results shown in the following section present corrected percentages of female and male researchers using Equations (2) and (3).

III. Case Study

For the 19 representative conferences of computer science selected for our analysis as shown in Table 1, we extracted from the DBLP Computer Science Bibliography a total of 10,003 names of authors contributing to these conferences within the last three years (original population). The names were then classified by origin and gender using the NamSor Applied Onomastics. From the original population, 2,068 author names assigned to Hong Kong, China, Taiwan, Republic of Korea, Viet Nam, and Democratic People's Republic of Korea were removed due to infeasibility of automatic classification. We also removed those names that are not classifiable due to the usage of initials instead of the full first name (174 names), and names that are not strongly correlated to gender (43 names).

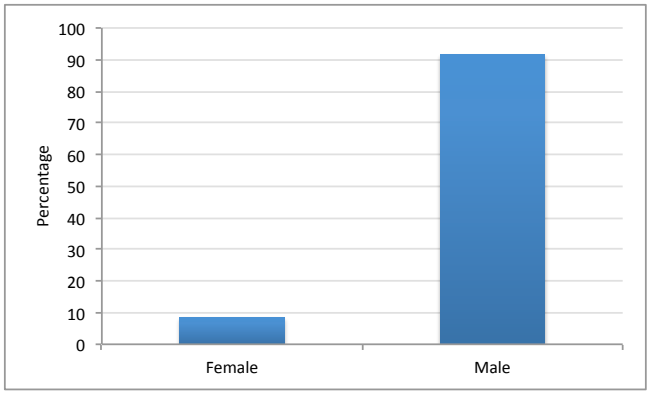


Figure 4: Final distribution of female and male names for 19 conferences in computer science and electrical engineering after removal of Asian and unclassified names, and correction using stochastic samples and applying Eqs. (2) and (3).

After applying the presented stochastic sampling of this population and subsequently applying the correction according to Eqs. (2) and (3) in Section II-C on the resulting basic population of 7,718 names, we could finally estimate that the percentage of women contributing to the 19 conferences within the last three years is, on average, below 10% (Figure 4).

Our approach now allows us to have a closer look at the proportion of scientifically active women in different individual conferences, and thus areas of computer science and not only to calculate the overall proportion of women in computer science as a whole. To illustrate the percentage of female authors in individual conferences, we picked out three of them: The ACM Symposium on Operating Systems Principles (SOSP), the Design, Automation and Test in Europe (DATE), and the International Conference on Applied Cryptography and Network

Security (ACNS). The percentage of female authors varies here between 5.5 percent for the SOSP and 10.4 percent for the ACNS conference. For the DATE conference, the percentage of female authors amounts to an average value of 8.3 percent (Figure 5).

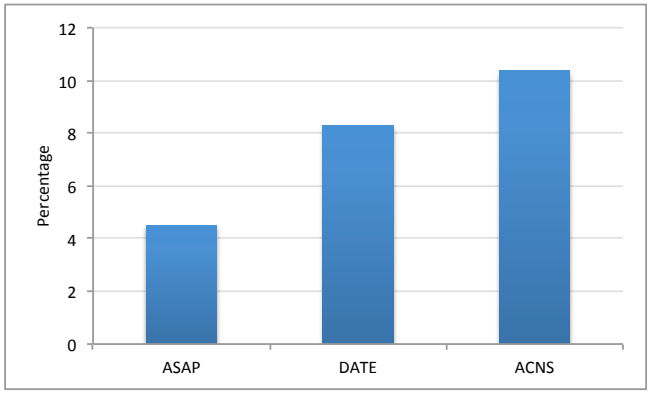


Figure 5: Participation of women in the conference with the highest (ACNS), the lowest (SOSP), and an average percentage of women (DATE).

A closer look at the participation of women in all 19 conferences finally reveals a slightly asymmetrical distribution (Figure 6 and Table IV). Only one of the investigated conferences has a percentage of female authors above 10 percent. Half of the conferences have a proportion of female authors below 8.1 percent, and 4 conferences have a proportion of female authors below 6.7 percent.

IV. Discussion

In this work, we presented a novel method to estimate the proportion of scientifically active woman in the specific scientific field of computer science. In contrast to previous studies that refer to limited data records, our method provides a more general approach with reduced limitations:
(1) We make sure to take only scientifically active researchers into account. Therefore, we decided to generate our population from the author lists of conference proceedings, assuming that scientifically active researchers publish their findings on conferences[2]. Besides postdoctoral and industrial researchers the examined group includes professors, senior lecturers and doctoral students. By considering only the databases of recent years, this approach allows us to exclude researchers that are not scientifically active anymore, for example, due to a change in their scientific field or job. Also,

[2]Conferences and the appropriate conference proceedings are the common publication medium in Computer Science and have a much higher impact than journal papers.

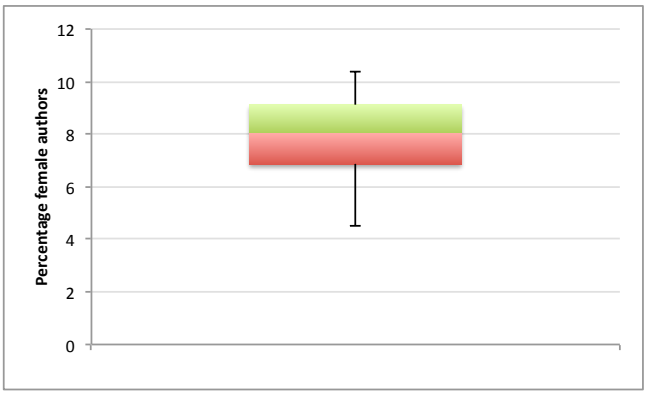


Figure 6: Box-and-Whisker chart illustrating the active participation of women in the conferences under consideration. Half of the conferences have a proportion of woman between 6.9 and 9.1 percent (1st and 3rd quartile) with the median at 8.1 percent. The conference with the most female authors has a proportion of women of 10.4 percent, the conference with the fewest female authors has a proportion of women of 4.5 percent.

Table IV

Conference	Percentage of female authors
ACNS	10.4
ARCS	7.8
ASAP	4.5
ASP-DAC	8.1
ASPLOS	7.3
CASES	6.1
CC	7.3
CODES+ISSS	9.1
DAC	8.8
DATE	8.3
Euro-Par	9.3
Eurosys	6.6
PARCO	8.6
SOSP	5.5
USENIX	7.5
VEE	9.3
DASIP	8.1
Humanoids	9.6
NDSS	6.9

researchers active in administration or management are omitted, as well as students at the beginning of their studies. In other words, only the relevant researchers are gathered by our method, non-relevant ones are excluded.
(2) We collect the data of our population independent of the university degree of the authors. This leads to a complete list of scientifically successful researchers, from doctoral students to full professors. Even career changer and employees without an academic degree like technicians or qualified IT specialist are included, as long as they successfully contribute to relevant conferences.
(3) We ensure to generate our population independent of the origin of the authors. On the selected international conferences you can find conference delegates from all over the world. As expected, we found author names from 89 different onomastic classes on our list, reflecting the likely country of origin of the authors. Our approach also provides the possibility to generate a population of authors only for national conferences or for individual conferences (Figure 5).

In contrast to previous studies searching for female scientists in the entire field of science or a discipline such as computer science, our approach makes it possible to further focus the analysis to a single conference or a set of representative conferences for a specific scientific field. For the case study presented here, we examined representative conferences suggested by the researchers of the CRC/Transregio 89 *Invasive Computing*, which covers computer engineering, operating systems, programming languages, security and the field of application including robotics and high-performance computing. By the selection of conferences, it would be possible to investigate also other scientific fields or to further limit the scientific area (e.g. to operating systems or computer security).

Despite these advantages of the method, we are not able to directly extract the gender or origin of the authors from the DBLP Computer Science Bibliography, one reason being that DBLP does currently not list these properties. In contrast to other studies, we were able to automatically determine the gender of the authors by NamSor Applied Onomastics. Yet, after testing the accuracy of this fully automatic classification on random samples from the group of men and women, we found out that although only 1 man was wrongly classified. 27 percent of women were men instead. A more thorough inspection indicated that 25.3 percent of wrongly classified women were from India. These differences in accuracy between men and women through verification by random sampling is not explained by NamSor Gender API. Indeed, they do not provide any information about classification of names from India. To take the wrong classifications into account, we determined corrective factors as explained in Section II-C.

The most significant disadvantage and a potential source of error of our approach is the removal of names classified as Asian names (see Section II-B). The excluded group comprises a total of 2,068 names, which amounts to 20.7 percent of all names in the original population obtained from the DBLP Computer Science Bibliography. The removal of these names may distort the results. However, there is no evidence so far that the proportion of women in the group of removed Asian names is significantly higher than in the investigated group. In fact, several studies on women in the STEM disciplines in Asia indicate that the proportion of female students is even lower than in other parts of the world ([14], [16]). For the approach introduced in this study, there was no possibility to determine the gender on the basis of an Asian name, as explained in detail in Section II-B. The use of the *Chinese Name Gender Guesser* [3] or other software platforms was not taken into consideration because these take the traditional Chinese characters of the name to classify the gender.

For our analysis, we also removed 217 additional names of unknown gender (see Section II-B), due to missing information. For example, 173 authors submitted only a single character as first name. There is obviously no way to determine the gender by one letter. However, there is no evidence that there is a disproportionate percentage of women in this group. These names reflect 2.2 percent of the entire population and were therefore neglected.
Another assumption taken in this study is the internet presence of the authors supposed for the estimation of the correction factors (see Section

II-C). This assumption, however, turned out not to be critical since the percentages of authors not found on the internet are in the same range for female and male authors.

In conclusion, we are presenting a novel method to capture and classify female scientists that are a) currently active in research and b) in each specific field of computer science. This group includes those female scientists successfully publishing their research findings in peer-reviewed publication media and, thus, having an impact on their scientific community. The data are collected regardless of the university degree and irrespective of whether the scientist is employed at university or industry. The data provided by the presented method are closing the gap of postdoctoral researchers in industry and university existing in many other surveys of women in science. Furthermore, the method allows to estimate the number of female candidates suitable for recruiting them as high-potential postdocs or professors.

This work was supported by the German Research Foundation (DFG) as part of the Transregional Collaborative Research Centre "Invasive Computing" (SFB/TR 89).

References

[1] Carsenat, Elian. *What's the Gender Gap in the European Union Whoiswho?* 2014. URL: http://blog.namsor.com/2014/09/09/whats-the-gender-gap-in-the-european-union-whoiswho/.

[2] Carsenat, Elian. *Cannes2015, Mind the gender map.* 2015. URL: http://blog.namsor.com/?s=cannes.

[3] Chinese Name Gender Guesser. *Chinese Name Gender Guesser.* URL: http://www.chinese-tools.com/tools/gender-guesser.html.

[4] dblp. *The DBLP computer science bibliography.* URL: http://dblp.uni-trier.de/db/.

[5] Linkedin. *Linkedin.* URL: https://www.linkedin.com.

[6] Marin, Gabriela, Barrantes E. Gabriela, Chavarria, Silvia. *Are women becoming extinct in the Computer Science and Informatics Program?* Jan. 2008.

[7] NamSor Gender API. 2015. URL: http://blog.namsor.com/api/.

[8] NamSor Origin API. 2015. URL: http://blog.namsor.com/name-recognition-software/.

[9] People's Daily. *Chinese surname shortage sparks rethink.* 2007. URL: http://en.people.cn/200706/19/eng20070619_385661.html.

[10] Ranga Marina, Gupta Namrata, Etzkowitz Henry. *Gender Effects in Research Funding.* 2012.

[11] Researchgate. *LinkedIn.* URL: https://www.researchgate.net.

[12] *She Figures 2012 – Gender in Research and Innovation.* European Commission, 2012. DOI: 10.2777/38520.

[13] Shokhenmayer, Evgeny Carsenat, Elian. *Onomastics to measure cultural bias in medical research sing scientists personal name.* 2014. URL: http://blog.namsor.com/2014/08/28/onomastics-to-measure-cultural-bias-in-medical-research/#_ftnref1/.

[14] *Studentinnenanteile in Mathematik, Naturwissenschaften und Informatik sowie Ingenieurwissenschaften im internationalen Vergleich, 2012.* CEWS - Center of Excellence Woman and Science, 2012. URL: http://www.gesis.org/cews/informationsangebote/statistiken/thematische-suche/detailanzeige/article/studentinnenanteile-in-mathematik-naturwissenschaften-und-informatik-sowie-ingenieurwissenschaften-im-internationalen-vergleich-2012/.

[15] Vichnevskaia, Tatiana. *Applying Onomastics to Scientometrics.* 2015. URL: https://inserm.academia.edu/taniavichnevskaia.

[16] *Women in Science and Technology in Asia.* AASSA - The Association of Academies and Societies of Sciences in Asia, Sept. 2015. URL: http://www.interacademies.net/Publications/28012.aspx.

[17] *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2015.* Na-

tional Science Foundation, National Center for Science and Engineering Statistics, 2015. URL: http://www.nsf.gov/statistics/wmpd/.