Scientometrics (2015) 103:85–100
DOI 10.1007/s11192-014-1520-3

The evolution of female authorship in computing research

José María Cavero · Belén Vela · Paloma Cáceres · Carlos Cuesta · Almudena Sierra-Alonso

Received: 16 June 2014 / Published online: 25 December 2014
© Akadémiai Kiadó, Budapest, Hungary 2014

Abstract In this paper we have conducted a study that covers computer science publications from 1936 to 2010 in order to analyze the evolution of women in computing research. We have considered the computing conferences and journals that are available in the digital bibliography and library project database, which contains more than 1.5 million papers and more than 4 million authorships, corresponding to about 4,000 journals, conferences and workshops. We analyze the participation of women as the authors of publications, productivity and its relationship with the average research life of women in comparison to that of men, the gender distribution of conference and journal authorships depending on different computer science topics, and authors' behavior as regards collaborating with one gender and/or the other. We also detail the method used to obtain and validate the data. The results of our study have led us to some interesting conclusions concerning various aspects of the evolution of female authorship in computing research.

Keywords Gender study · Research publications · Computer science · DBLP database

J. M. Cavero · B. Vela (✉) · P. Cáceres · C. Cuesta · A. Sierra-Alonso
Research Group VorTIC3, Computer Science and Statistics Department, Universidad Rey Juan Carlos, C/Tulipán s/n, 28933 Móstoles, Madrid, Spain
e-mail: belen.vela@urjc.es

J. M. Cavero
e-mail: josemaria.cavero@urjc.es

P. Cáceres
e-mail: paloma.caceres@urjc.es

C. Cuesta
e-mail: carlos.cuesta@urjc.es

A. Sierra-Alonso
e-mail: almudena.sierra@urjc.es

Introduction

What percentage of computer science researchers are women, and how has this percentage evolved since the beginnings of this discipline? Are there any differences between men and women's productivity? Are there any differences between men and women as regards their duration as researchers, which may influence their productivity? Are women more likely to publish in certain computing areas than in others? What is the behavior of authors like when collaborating with one gender and/or the other?

In this paper we have attempted to answer these questions by conducting a study of computer science publications from 1936 to 2010 in order to analyze the evolution of women in computing research. We have considered the computing conferences and journals that are available in the digital bibliography and library project (DBLP) database (http://www.informatik.uni-trier.de/~ley/db/), which includes more than 1.5 million papers (starting from 12 papers in 1936 to about 200,000 in 2010), more than 4 million authorships (more than 900,000 different researchers), corresponding to about 1,000 different journals and 3,000 different conferences and workshops.

In what remains of the paper we consider a researcher to be a person who carries out research and who publishes his or her results in a paper, while authorship is considered as the instance of a researcher being the author of a paper. Therefore, one researcher usually corresponds to several authorships (as many papers as he or she writes).

Several works appertaining to other disciplines have investigated gender differences in scientific publications. Sax et al. (2002) and Xie and Shauman (1998) analyze the gap in research productivity between men and women researchers from universities in the United States of America (USA) and that of US researchers in four large nationally representative cross-sectional surveys of postsecondary faculties in 1969, 1973, 1988, and 1993. In Mauleón and Bordons (2006), both the research productivity and impact and publication habits of materials science researchers are analyzed by gender. Aksnes et al. (2011) assume the well-established conclusion that "female scientists tend to publish fewer publications than do their male colleagues" in order to analyze in their work whether similar gender differences can also be found in terms of citations. Some gender research can also be found in disciplines closer to computer science. Gallivan and Benbunan-Fich (2006) studied the research productivity of the top researchers in information systems journals. They compared their data with previous studies and with an estimated population of information system researchers, and found that 17 % of the top 251 researchers were women.

Some gender studies exist in computer science, although they do not focus on the publication domain. There are related data in the National Center for Education Statistics (2011): the number of female PhD graduates in computer science in the USA from 1969 to 2010 was 17.22 %. Moreover, there are fewer women than men in computer science in higher education (Papastergiou 2008) with a few exceptions (Gharibyan and Gunsaulus 2006; Othman and Latih 2006). Ceci and Williams (2011) analyze the causes of women's underrepresentation in science, concluding that discrimination is not the cause of this in some fields of science. Women's participation in computer science has been compared to a shrinking pipeline (Camp 1997), in which the ratio of women decreases as regards the amount of female students in comparison to the amount of women who hold positions in academia.

We have conducted our study because, despite the fact that several works investigate gender differences in scientific publications, to the best of our knowledge only the paper published by McGrath Cohoon et al. (2011) makes a study of the historical evolution of female participation in the computing discipline. These authors, however, only analyze the

evolution of female authors in ACM computing conferences from 1966 to 2009. In Sect. 4 we compare their results with ours.

The results of our study have led us to some interesting conclusions concerning various aspects of the evolution of female authorship in computing research: female authorships have increased from about 3 % to more than 16 % in the last 50 years; the number of papers by women is less than that of men, but this seems to result from the fact that the average research life of women is shorter than that of men; some variations in female authorship exist if we analyze women's participation in different areas of computing (they tend to do more research in Human–Computer Interaction and less in areas such as Computer Vision and Pattern Recognition); the real percentage of papers authored only by women exceeds the expected percentage.

In the following sections, we first detail the method used to obtain and prepare the data from the DBLP database. We then present the results related to the role of female authors. We specifically analyze their participation as the authors of publications that appear in computing conferences and journals, their productivity, the average research life of women with regard to that of men, followed by the way in which women publish depending on different computer science topics and their behavior as regards collaborating with one gender and/or the other. We then show the validation of the data, and finally we present our conclusions.

Obtaining and preparing the data

We decided to use the data from the DBLP database to define our population (researchers and papers) as it is the most complete and open access repository of computer science publications, although we are aware of the fact that the DBLP database is not a complete and unbiased source as it has different coverage for different Computer Science Areas (Wainer et al. 2013).

The data used in our study was obtained from an XML file (more than 1 GB) provided by the DBLP. This complex XML file contains all the biographic records including, among other things, information concerning: different categories of publications (journals, conferences, books, series, etc.), authors and editors, and publications (identifier, title, volume, issue, pages, etc.). We then created relational tables which were directly transformed from the XML file, and we subsequently modified the relational database structure in order to store additional information not included in the original data, such as the gender of the authors, or the classification of certain journals according to the SJR (SCImago Journal & Country Rank, SJR 2010) Computer Science categories and JCR (Journal Citation Reports, JCR 2010).

When performing data cleansing it was also necessary to detect and correct certain issues in the data from the XML file. One of these issues was the fact that DBLP stores information concerning the case of two authors being the same person. We have modified this information by unifying such entries into one. Moreover, as there are some redundant and useless data for our purpose, it has been necessary to clean them in order to leave only accurate and consistent information in the database.

As the DBLP does not contain the gender of the researchers, it was then necessary to identify the gender of each of the authors, and we have therefore created a database of names for gender identification obtained from several sources. Our main source was the US census of 1990 (http://www.census.gov/genealogy/www/data/1990surnames/names_files.html), which contains over 5,000 names, including the number and percentage of males and females with that name. Although the data from the US census are fairly complete, we

have completed them with some data from a Spanish census and other web sources, which has allowed us to identify the gender of a name in a non-ambiguous manner.

Nevertheless, tens of thousands of researchers' genders remained unknown (by unknown we mean that the gender of a researcher who authors a specific paper is not known). 35.46 % (1,443,113) of all authorships corresponds to researchers of unknown gender. Of that, 215,565 authorships (5.30 % of the total) correspond to researchers whose name is just an initial, For example, more than 22,000 papers correspond to researchers whose names appear in the database with solely the initial 'M.'. Almost 300,000 authorships correspond to researchers with ambiguous names, signifying that they cannot be used to identify the gender with certainty (for example, Chris, Alex, Jean, …).

Another important aspect is that in last few years the presence of Asian (mainly Chinese) researchers has grown and our database of names (including the US census data) identifies the gender of very few Chinese researchers' names. However, this issue does not greatly influence the result of our study as it covers the time period from 1936 to 2010 and the increase in the influence of Chinese researchers in Computer Science has taken place in the last few years, especially since approximately 2007, as can be seen in the country rank of the SJR website (http://www.scimagojr.com/countryrank.php).

Various solutions for the treatment of ambiguous names exist, such as using a method to estimate people's gender or assuming that the gender distribution of the unambiguous names is the same as that of ambiguous names. In an attempt to "estimate" the gender of people with ambiguous names we used the same method as that used by US census distribution to predict the gender of a name. For example, according to the US census, 75,219 US female citizens are called Shawn, and 57,973 male citizens are also called Shawn. We transformed this data into percentages in order to obtain the probability of a person called Shawn being male or female. However, we eventually decided not to estimate gender and to assume that the gender distribution of the unambiguous names is the same as that of ambiguous names, since during the validation of this method (see Sect. 4) we obtained better results when using only unambiguous names.

As mentioned previously, in our study we have considered more than 1.5 million papers and more than 4 million authorships who participated in computing conferences and journals from 1936 to 2010. We identified the gender of 2.6 million authorships with certainty (corresponding to more than 500,000 different researchers) based on their names and using the aforementioned database of names.

The quality of our results obviously depends not only on the correct gender having been assigned to the authors but also on the quality of the data provided by the DBLP. The problem of data quality in the DBLP is explained in Ley and Reuther (2006) and Ley (2009). These authors describe some of the algorithmic solutions used to detect errors in the identification of researchers. For example, they check whether people with a "similar" name who have the distance of two in the coauthor graph may be the same person. Although they attempt to identify the people behind the research papers and to treat synonymous and homonymous names as precisely as possible, it is obvious that some errors may persist, in spite of the procedures established. However, since in this paper the same data is used throughout the entire period of time studied, we consider the results obtained to be valid.

Publishing behavior of female researchers in computer science since its beginnings

The participation of women in computer science research has grown since the beginnings of the discipline. Figure 1 summarizes the evolution of female participation in computer

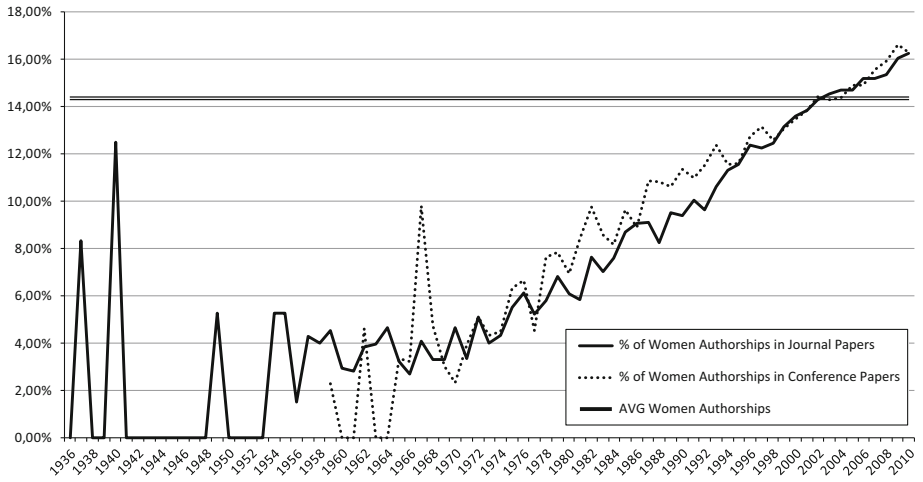

Fig. 1 Evolution of female authorships in computer science research

science research, considering the data related to the computer science papers available in the DBLP database from 1936 to 2010. This shows that in 1966, <3 % of the authorships on computing published were written by women, as opposed to about 16.3 % in 2010. The graphs for journals and conferences are shown separately and, as can be seen, there have been no significant differences since the late 1990s. The quantity of journal papers available during the 1930s, 1940s and 1950s, and the quantity of conference papers available during the 1960s are small, and great variations can therefore be observed during these years.

Of the 2.6 million authorships considered (we have considered only those authorships whose gender has been identified), 14.33 % are women. This average value appears in Fig. 1 as a double horizontal line. This value is very close to the percentage regarding female authorships in the last few years. The reason for this is that the quantity of papers and authors has dramatically increased during the last few decades and this has therefore had a great influence on the total. These 2.6 million authorships correspond to more than 500,000 different researchers, 18.88 % of whom are women. These data may suggest that women produce, on average, less than expected, because women represent the 18.88 % of the researchers but they correspond to only 14.33 % of the authorships.

We shall now attempt to quantify this difference in terms of productivity. There are many different approaches to productivity measurement. In this paper we measure productivity in terms of quantity, although for a more complete definition we should consider not only quantity but also quality. This obviously implies various difficulties, as it involves comparing journals with conferences of various categories and amounts of prestige, and it does not consider other aspects such as impact factors. However, we consider it a valid criterion, as the same measurement is used for the whole period of time. Moreover, we have also analyzed productivity by considering only those journals listed in the JCR (2010). We have considered the data starting from 1953, as the data regarding JCR publications are available from that year in the DBLP database. The result obtained after comparing the female productivity in JCR journals with that of all DBLP journals is very similar, as can be observed in Fig. 2.

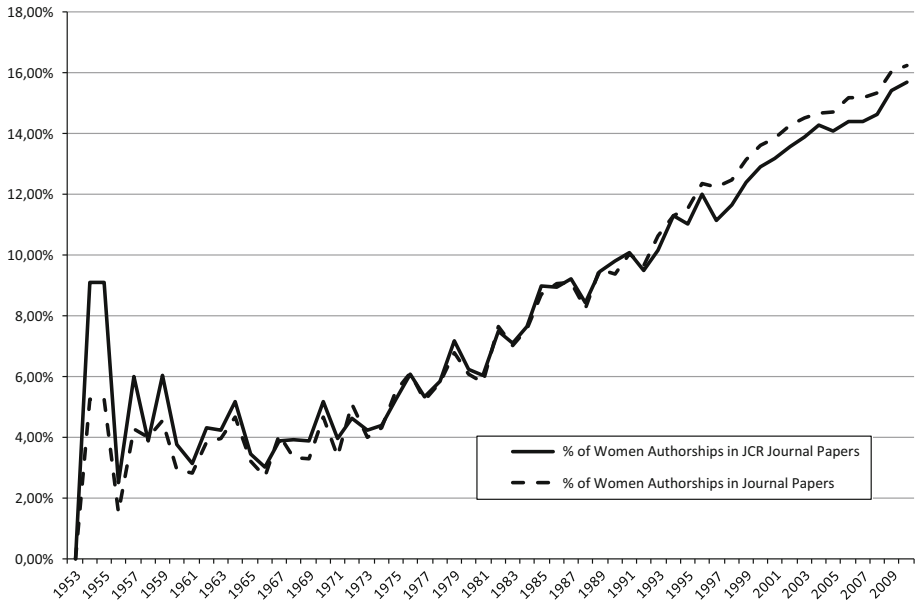

Fig. 2 Female authorships in DBLP journals versus JCR journals

In an attempt to analyze whether the difference in the number of papers published by men and women could be explained by a difference in the duration of their research lives, in Table 1 we have compared the productivity of researchers with their average research lives.

The first column of Table 1 indicates the minimum number of papers authored by the researchers considered in each row (minimum number of papers). The first row therefore shows the information relating to researchers with at least one paper, that is, the whole population. The second row shows the information relating to researchers with at least two papers (which is a subset of the previous row), and the last row shows the information relating to researchers with at least 30 papers. The next three columns represent the average number of papers per researcher: average number of papers authored by men (men), average number of papers authored by women (women) and the difference in productivity between men and women as a percentage (% diff). For example, of the researchers with at least 10 papers, each man produces 28.92 papers, and this represents 10.52 % more papers than each woman. The following three columns include the information concerning the average research life of men and women depending on the number of papers they have published: average research life of men (men), average research life of women (women) and percentage of the difference between men and women's research lives (% diff). We define research life as the difference between the year of the last paper and the year of the first paper authored by a researcher. For example, the average research life of a female researcher who has published at least 5 papers is 10.22 years. Finally, an additional column (% of women) shows the percentage of female researchers in that population.

As can be seen, the difference in productivity between men and women decreases as the minimum number of papers increases. Moreover, the difference in men and women's average research lives decreases as the minimum number of papers increases.

Table 1 Comparison of productivity of researchers and their research lives

Minimum number of papers	Average number of papers			Average research life			% of women
	Men	Women	% Diff	Men	Women	% Diff	
1	5.38	3.86	39.19	4.51	3.33	35.16	18.88
2	9.34	7.48	24.91	7.69	6.28	22.40	16.40
3	12.51	10.46	19.57	9.46	8.01	18.15	15.39
4	15.29	13.10	16.71	10.71	9.23	15.98	14.80
5	17.84	15.60	14.38	11.68	10.22	14.33	14.31
10	28.92	26.17	10.52	14.83	13.36	11.05	13.29
15	38.48	34.81	10.54	16.77	15.23	10.12	12.99
20	46.88	42.99	9.06	18.10	16.59	9.09	12.49
25	54.62	50.64	7.86	19.07	17.57	8.54	12.07
30	61.82	57.31	7.87	19.82	18.34	8.07	11.91

Table 2 Percentage of women authorships (2001–2010) in SJR journals, classified by areas and ordered by percentage (in the case of the row total SJR, the total papers, % of papers and total journals columns are not the sum of the areas because many journals are included in more than one area)

Area	% of female authorships	Total papers	% of papers	Total journals
Computer Vision and Pattern Recognition	12.14	17,144	8,82	20
Hardware and Architecture	12.17	24,648	12,68	46
Signal Processing	12.31	14,647	7,53	17
Computer Graphics and Computer-Aided Design	13.51	26,060	13,40	39
Computer Networks and Communications	13.55	23,083	11,87	57
Computer Science, Artificial Intelligence	13.72	43,111	22,17	66
Software	14.32	33,831	17,40	71
Computer Science, Computational Theory and Mathematics	14.70	49,304	25,36	69
Total SJR	14.73	194,424	100,00	378
Information Systems	16.21	37,036	19,05	69
Computer Science Applications	17.21	43,003	22,12	71
Computer Science (Miscellaneous)	17.97	12,883	6,63	46
Human–Computer Interaction	19.51	3,166	1,63	12

The data presented in Table 1 allows us to conclude that the difference in the duration of men and women's lives as researchers could explain the difference in the number of papers published by men and women. In fact, if we consider researchers who have published at least 5 papers, the difference in the average research life and difference in productivity is nearly the same. As can be observed in the last column of Table 1, as the minimum number of papers published grows, the percentage of female researchers decreases.

These results are consistent with many studies that show that male scholars publish more than their female colleagues. Several works have studied the reasons behind this productivity puzzle (Cole and Zuckerman 1984). The consequence of most of these causes may be that women abandon their careers before men, or initiate their careers at a later age than men (see the recent work of Van Arensbergen et al. (2012) for a summary of these causes). The data shown in Table 1 quantify the effects of the causes and provide a quantitative explanation of the productivity difference between men and women.

However, in order to be really able to interpret this data, it would be very helpful to know the type of researcher we are talking about (professional researchers, staff, students, transitioning researchers, etc.). The available data allows us only to surmise the category of the researchers, so, for example, we can only surmise that professional researchers would be those who have published more papers.

We shall now analyze how women publish depending on the various computer science topics. This has been done by showing the classification of the journals according to the SJR used in the analysis. Table 2 shows the percentage of female authorships of journal papers depending on their classification in the SJR during the last 10 years (we only present the analysis of recent data and not their historical evolution). As Table 2 shows, Computer Vision and Pattern Recognition and Hardware and Architecture are the two areas with the lowest percentage of female authorships. 14.73 % of authorships that publish in SJR journals are female. The trend over time (not shown in this paper) shows that the percentage of female authorships who publish in SJR journals is very similar to that of female authorships who publish in the other journals. The area in which women participate most is that of Human–Computer Interaction.

In order to analyze what we term as authorships' 'collaborative behavior' (i.e., the behavior of the authors of the papers as regards collaborating with one gender and/or the other), we have grouped the papers into seven categories:

1) Only men: 555,067 papers (36.19 %), including single author papers. The gender of all the authors of these papers is known and male;
2) Only women: 40,359 papers (2.63 %), including single author papers. The gender of all the authors of these papers is known and female;

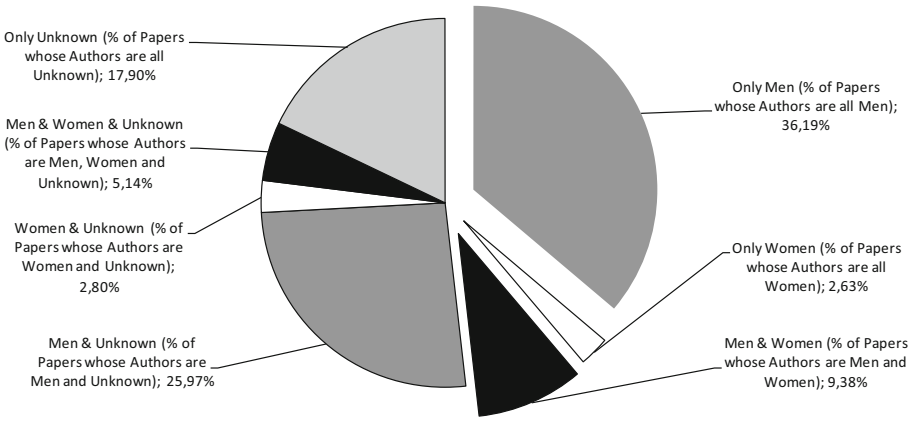

Fig. 3 Classification of papers according to different categories of gender collaboration

3) Only unknown: 274,527 papers (17.90 %), including single author papers. The gender of all the authors of these papers is unknown;
4) Men and women: 143,870 papers (9.38 %). The gender of all authors of these papers is known and there is at least one male and one female author;
5) Men and women and unknown: 78,864 papers (5.14 %). There is at least one male author, one female author and one author whose gender is unknown;
6) Men and unknown: 398,296 papers (25.97 %). There is at least one male author and at least one author whose gender is unknown (i.e., there are no known female authors); and
7) Women and unknown: 42,970 papers (2.80 %). There is at least one female author and at least one author whose gender is unknown (i.e., there are no known male authors).

Figure 3 shows the results of the grouping: only 2.63 % of papers are written solely by women versus 36.19 % of papers written exclusively by men. The percentage of papers written only by women might increase if we, albeit optimistically, assume that the unknown authors for women and unknown papers are female.

In order to address this issue, we shall hereafter consider only those categories in which the gender of all authors is known, that is, if the gender of at least one author of a paper is unknown, then that paper will be excluded.

Figure 4 shows the percentages of the three categories in which the gender of all authors is known: only men, only women and men and women. The result of the grouping is that only 5.46 % of papers are written solely by women versus 75.08 % of papers written exclusively by men, including single author papers.

Figure 5 shows the historical evolution of the percentage of papers in the categories included in Fig. 4. Of all the papers studied, the percentage of papers on which women and men collaborate (men and women category) has increased over time (26.03 % in 2010).

This set of papers without unknown authorships has been used as a starting point to calculate the percentage of papers on which only women collaborate (including the single author papers), and we have compared this with the expected percentage of papers written only by women. In order to obtain the expected number of collaborations among females, for each year we have taken into account the percentage of female authorships and the

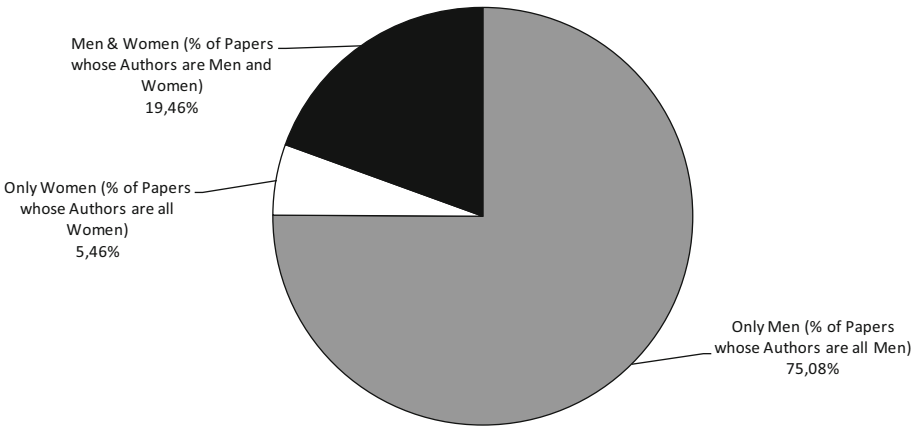

Fig. 4 Classification of papers according to the categories: only men, only women and men and women

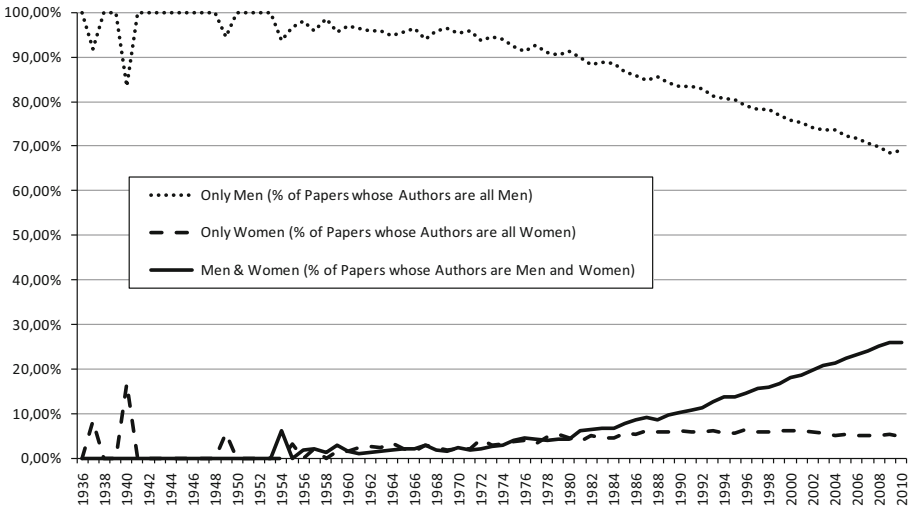

Fig. 5 Evolution of male/female collaboration according to the three categories: only men, only women and men and women

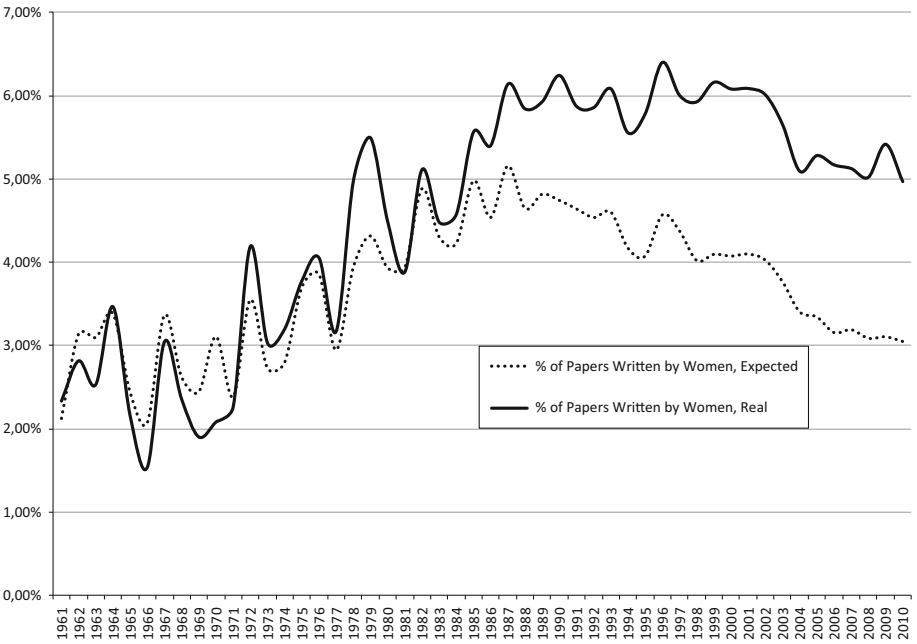

Fig. 6 Expected versus real percentage of papers written only by women

number of papers that are authored by a given number of researchers. The expected number of papers written only by women in a given year is therefore the sum of the expected number of papers written by one woman and the expected number of papers

written by two women and the expected number of papers written by three women and so on …That is, we have calculated the expected total number of papers written only by women in a given year y, TPW_y, in the following manner:

$$\mathrm{TPW}_y = \sum_{i=1}^{n_y} \mathrm{TP}_y(i) * (P_y)^i$$

$\mathrm{TP}_y(i)$: total number of papers in the year y with i authors, n_y: maximum number of authors per paper in the year y, P_y: percentage of female authorships in the year y

The Fig. 6 shows the percentage of papers authored by only women, expected versus real. As can be seen, both lines behave similarly as regards growth until the eighties (owing to the growing percentage of female authorships: most papers were authored by one person). Then, while the expected percentage fell, the real percentage still grew, or fell slightly. In our opinion, the reason for this behavior is strongly related to the average number of authorships per paper (Cavero et al. 2014). At the beginning of the eighties, the average number of authorships per paper was still 1.6. It then underwent a rapid growth: at the beginning of the nineties, the average number of authorships per paper was 2; in 2010, it was 3. Therefore, although the percentage of women authorships continued to grow, the rapid growth in the number of average authorships per paper caused the expected percentage of papers written by women to fall (as the average number of authorships per paper grows, it is less likely that all of them will be women). Nevertheless, the real percentage clearly exceeds the expected one. In our opinion, this indicates that women are more likely to collaborate with women than with men.

Validation of the method

We validated the method by comparing the results obtained with a manual gender identification in two different studies regarding authorships in software engineering (Vela et al. 2012) and information systems (Gallivan and Benbunan-Fich 2006) journals with the results obtained using our automatic method for the authors of the same journals. In both cases, our results were similar to those obtained using manual gender identification, as will be shown below.

For the first validation we used one of our previous works (Vela et al. 2012). We analyzed the authors of a set of the top 12 Software Engineering journals during 2007 and 2008, obtaining that 17.26 % of the authors were women. Our automatic method was then used to carry out the same analysis, but using only ten of these journals (TOSEM, EMSE, IEEE SW, TSE, IST, IJSEKE, JSME, JSS, RE, SQJ—the other two, IET SW and STVR, are not contained in the DBLP database), and obtained that 17.23 % of the authors are women.

For the second validation we used the work by Gallivan and Benbunan-Fich (2006), which addresses the percentage of women authors among the most prolific authors in a set of 12 Information Systems journals from 1999 to 2003, and concludes that 16.7 % of the most prolific authors (at least 3 papers) in Information Systems were women. As in the previous case, the same analysis was carried out, but considering only the subset of the journals available in the DBLP, in this case 9 of these 12 journals (EJIS, DSS, ISJ, IAM, ISR, JMIS, ORGSCI, MISQ, JSIS), and our method allowed us to obtain almost exactly the same results: 16.19 % of authors with at least 2 papers and 17.53 % of authors with at least 3 papers are women.

After carrying out these tests, we can therefore state that our gender identification method allows reliable conclusions to be obtained.

This result differs from the results of a paper published in the CACM in August, 2011 by McGrath Cohoon et al. (2011), which concluded that the percentage of female authors in ACM computing conference papers has increased from 7 % (1966) to 27 % (2009). According to our method, in 1966, <3 % of the authors of published computing papers were women, as opposed to about 16.4 % in 2009 and 16.3 % in 2010.

The differences between the aforementioned work and ours are summarized in the following paragraphs, along with an attempt to identify possible reasons for these differences.

Firstly, McGrath Cohoon et al. (2011) analyzed 86,000 papers from more than 3,000 ACM conferences (including the different editions of the same conference) from 1966 to 2009. They sought to identify the gender of paper authors based on author names, using a database of names. In order to identify the gender of "unknown" or "ambiguous" names, McGrath Cohoon et al. used a method that assessed the probability of a name being either male or female. In our opinion, the contradiction between their results and ours does indeed result from the fact that their estimation of "unknown" or "ambiguous" names overestimates the number of women publishing in ACM conference proceedings. One reason for this may be that the popularity of using certain ambiguous names for men or women varies over time (for example, according to the US Social Security Administration (http://www.ssa.gov/OACT/babynames/), people called Riley are probably men if they were born in 1995, but probably women if they were born in 2005). What is more, the gender associated with a particular name may vary according to the country that the person is from (authors called Andrea are probably women, except if their nationality is Italian, in which case they are probably men). Another reason may be that the authors have made certain assumptions as regards the distribution of the researchers with ambiguous names.

McGrath Cohoon et al. claimed that they identified the gender of 90 % of 356,703 authorships (using a method which assesses the probability of the name being either male or female for the "unknown" or "ambiguous" names). Our approach identified the gender of more than 2.6 million authorships, not taking into account almost 300,000 ambiguous and 1.1 million unknown names (with 220,000 being limited solely to initials).

Finally, McGrath Cohoon et al. compared their results with statistics regarding the gender of Ph.D. holders, concluding that the productivity of women is higher than that of men (since women's productivity was much higher than the percentage of female Ph.D holders). Recognizing that this result somehow contradicted the established conclusions of many studies that show that male scholars publish more than their female colleagues (Abramo et al. 2009; Fox and Mohapatra 2007; Fox 2005), they proposed possible explanations for the contradiction, including the theory that "men and women might tend to publish in different venues, with women overrepresented at ACM conferences compared to journals, IEEE, and other non-ACM computing conferences." However, according to our data, female authorships in the ACM and other conferences follow a similar trend, as can be seen in Fig. 7. We have considered the same time slice as that of McGrath Cohoon et al. (2011), that is, from 1969 to 2010. In our opinion the contradiction in McGrath Cohoon et al.'s results is owing to the fact that their estimation of the "unknown" or "ambiguous" names overestimates the number of women publishing in ACM conference proceedings.

In order to compare our results with those of McGrath Cohoon et al. (2011), we have created a similar chart to that presented in the aforementioned work. In Fig. 8 we have considered the same period of time and we have compared the percentage of female PhD

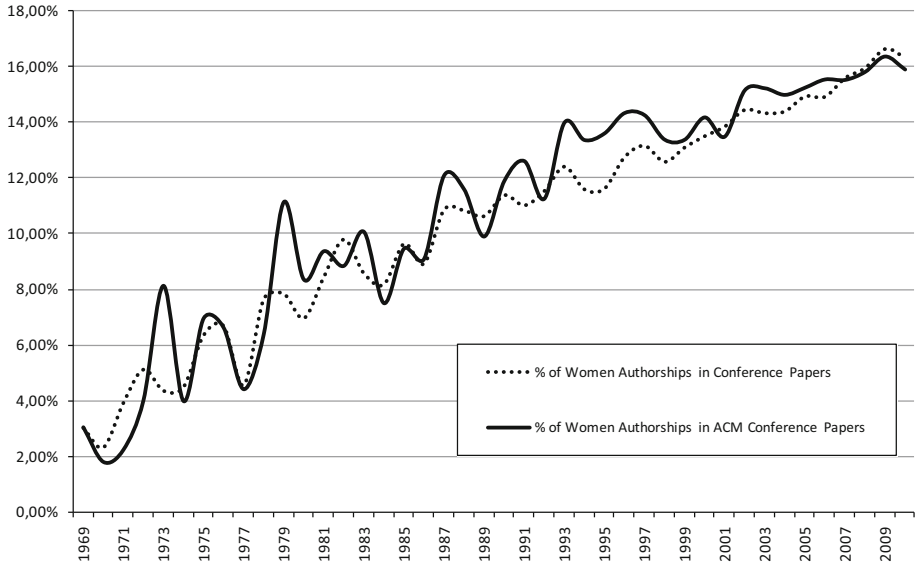

Fig. 7 Percentage of women authorships in ACM conferences versus in all conferences

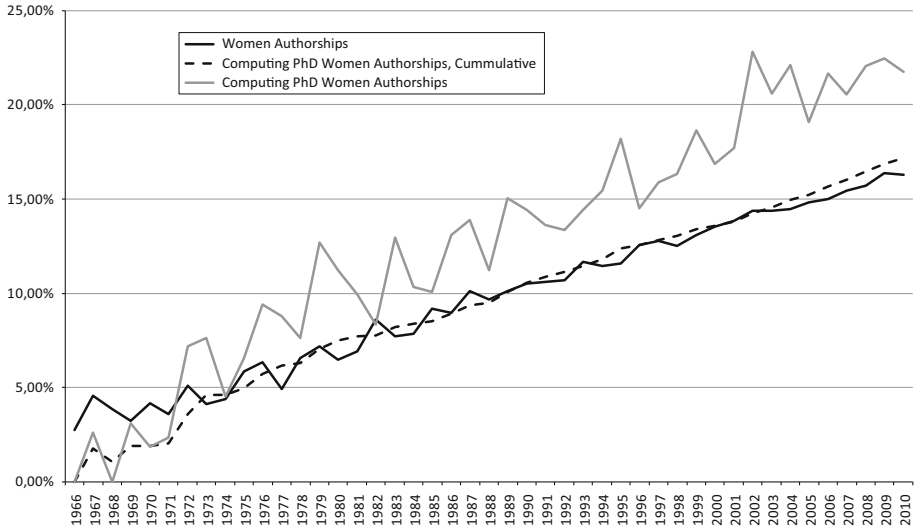

Fig. 8 Percentage of women authorships, of women PhD holders and of cumulative women PhD holders since 1966

graduates, the percentage of cumulative female PhD holders in the US and the percentage of female authorships worldwide. We are aware that Fig. 8 has some drawbacks: it mixes authorship with researchers (authorships of PHD dissertations are really researchers because everybody only authors a single PhD dissertation); it mixes worldwide data with that of the US; and, finally, although the objective of the cumulative percentage of women

PhD holders is to represent the total proportion of female researchers (considering that PhD holders continue to be active researchers for 30–40 years) a more profound study of how many years should be considered in this cumulative number is necessary. In spite of all these drawbacks, we have included the graph for comparison purposes. The conclusion of our comparison was that our results coincide almost exactly with the cumulative quantity of women PhD holders in Computer Science in the US (National Center for Education Statistics 2011), whereas in the work of (2011) the percentage of female authorships is much higher than the percentage of female cumulative PhD graduates (in 2009 it was approximately 27 vs. 17 %).

Conclusion

In spite of the significant involvement of women in early computer science history, the discipline itself has been archetypically characterized as a male-only field. While this is also true for many other technical disciplines, recent studies show that the situation is even more critical in computing-related areas (Guzdial 2012).

In the last 50 years the percentage of women scholars in computer sciences has grown from 3 % to more than 16 %. The results of our study are consistent with many results and findings, although they are lower than those obtained by McGrath Cohoon et al. (2011). In our opinion, this contradiction between their results and ours is indeed owing to the fact that McGrath Cohoon et al.'s estimation of "unknown" or "ambiguous" names overestimates the number of women publishing in ACM conference proceedings.

The participation of women as researchers in computer science has grown over the last 50 years (1960–2010) at a compound annual growth rate of about 3.5 %. Although the data could lead us to believe than this growth may continue in the future, some recent symptoms may alter this trend: the percentage of female computer science Bachelor's degree holders in the US has decreased over the last few years (Guzdial 2012). If this trend spreads to the percentage of women computer science PhD holders, it will probably affect the percentage of female computer science researchers. Some of the traits associated with traditional male-dominated technical areas can certainly be applied here: some of them can be simply considered to be a consequence of sheer numbers (the majority of males in the classroom is in itself a reason for some females to think twice about studying a degree of this type), while some others are probably related to the self-perpetuation of certain stereotypes. Both categories include drawbacks such as the female perception of isolation, the slow growth of the gender ratio, smaller revenues and turnover rates, or the wider influence of male senior researchers (the classic' old boys' network').

The well-known study by Margolis and Fisher (2002) investigated the reasons for this gender gap within the computer science area, described by the authors as a "male clubhouse", and related it specifically to an early educational and social perception. They concluded that in a context of educational reform, the figures may significantly evolve: in the case of the Carnegie Mellon University, the percentage of females increased from 7 to 42 % in just 5 years. However, 10 years later the number of women in computing has not significantly increased, particularly in the US (Guzdial 2012). Some recent studies indicate a clear influence of the representation of genders in the Media (Cheryan et al. 2013) on the female acceptance of computing, showing that women's perceptions after an actual programming course changed significantly from previous pre-conceptions. The early presence of other scientific disciplines (such as Math or Chemistry) in primary education would thus explain their greater popularity among women. This should therefore also benefit from

recent campaigns advocating the early introduction of computing in schools, particularly in the UK (Berry 2013).

Another conclusion of our study is related to the difference in productivity between men and women. This confirms some established conclusions which show that men publish more than women, but this seems to result from the fact that the average research life of men is longer. The reason for this may be similar to the reasons given by Ceci and Williams (2011) to explain the underrepresentation of women in science, i.e., it is primarily owing to "factors surrounding family formation and childrearing, gendered expectations, lifestyle choices and career preferences". It would be necessary to carry out an analysis of the choices and pressures that women confront as regards their research life duration. Another interesting issue to consider is the possible existence of gaps of time in the research life, as we simply consider the difference between the first and the last paper of a researcher, and to analyze whether there are different patterns according to the gender.

Our study also shows some small preferences in the publishing habits of women depending on the various computer science topics. The areas of Computer Vision and Pattern Recognition and Hardware and Architecture are the two with the lowest percentage of female researchers, while Human–Computer Interaction is the area in which more women participate. These results of preferences in the publishing habits of women in the different computer science areas are consistent with those of other studies. For example, the work of McGrath Cohoon et al. (2011) shows that conferences focusing on human factors and on documentation are associated with a greater proportion of women authors and they suggest that these finding might be owing to "the hypotheses that alignment with gender stereotypes predicts the extent of women authorship".

With regard to the behavior of authors when collaborating with one gender and/or the other, the percentage of papers on which women and men collaborate has increased over time (26.03 % in 2010). In our opinion, there are two reasons for this: first, the percentage of women researchers is increasing with time and second, the average number of authors per paper has also grown. Cases in which all the authors of one paper are of the same gender are therefore less frequent. Moreover, the percentage of papers on which only women collaborate exceeds the expected percentage from the eighties. In our opinion, this indicates that women are more likely to collaborate with women than expected.

We are currently carrying out a study to analyze the publication pattern and academic context of the 100 most productive researchers. For this study, we are retrieving data, in most cases manually, from various Web sources (personal Websites, Websites of institutions, etc.).

Acknowledgments This research has been carried out in the framework of the following project: Co-Mobility (TIN2012-31104) financed by the Spanish Ministry of Economy and Competitiveness.

References

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Gender differences in research productivity: A bibliometric analysis of the Italian academic system. *Scientometrics, 79*(3), 517–539.

Aksnes, D. W., Rørstad, K., Piro, F. N., & Sivertsen, G. (2011). Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of American Society for Information Science and Technology, 62*(4), 628–636.

Berry, M (2013). Computing in the national curriculum: A guide for primary teachers, Naace & Computing at School.

Camp, T. (1997). The incredible shrinking pipeline. *Communications of the ACM, 40*(10), 103–110.

Cavero, J. M., Vela, B., & Cáceres, P. (2014). Computer science research: More production, less productivity. *Scientometrics, 98*(3), 2103–2111.

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of the United States of America (PNAS), 108*(8), 3157–3162.

Cheryan, S., Plaut, V. C., Handron, C., & Hudson, L. (2013). The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women. *Sex Roles: A Journal of Research, 69*(1–2), 58–71.

Cole, J. R., & Zuckerman, H. (1984). The productivity puzzle: Persistence and change in patterns of publication of men and women scientists. In P. Maehr & M. W. Steinkmap (Eds.), *Advances in motivation and achievement* (pp. 217–258). Greenwich: JAI Press.

Fox, M. F. (2005). Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science, 35*(1), 131–150.

Fox, M. F., & Mohapatra, S. (2007). Social-organizational characteristics of work and publication productivity among academic scientists in doctoral-granting departments. *Journal of Higher Education, 78*(5), 543–571.

Gallivan, M. J. & Benbunan-Fich, R. (2006). Examining the relationship between gender and the research productivity of IS faculty. Proceedings of the ACM SIGMIS conference on computer personnel research ACM New York. NY, USA: ACM Press. pp. 103–113.

Gharibyan, H. & Gunsaulus, S. (2006). Gender gap in computer science does not exist in one former Soviet republic: Results of a study. In Proceedings of the 2006 ACM ITiCSE conference on innovation and technology in computer science education. ACM New York. NY, USA: ACM Press. pp. 222–226.

Guzdial, M. (2012). U.S. Women in computing: Why isn't it getting better? Blogs of the communications of the ACM. http://m.cacm.acm.org/blogs/blog-cacm/149681-us-women-in-computing-why-isnt-it-getting-better/fulltext. Accessed 2014.

JCR. (2010). *Journal citation reports®*. Philadelphia, PA: Thomson Reuters.

Ley, M. (2009). DBLP: some lessons learned. *PVLDB, 2*(2), 1493–1500.

Ley, M., & Reuther, P. (2006). Maintaining an online bibliographical database: The problem of data quality. Revue des Nouvelles Technologies de l'Information RNTI-E-6. *Cépaduès-Éditions, 2006*, 5–10.

Margolis, J., & Fisher, A. (2002). *Unlocking the clubhouse: Women in computing*. Cambridge, MA: MIT Press.

Mauleón, E., & Bordons, M. (2006). Productivity, impact and publication habits by gender in the area of materials science. *Scientometrics, 66*(1), 199–218.

McGrath Cohoon, J., Nigai, S., & Kaye, J. (2011). Gender and computing conference papers. *Communications of the ACM, 54*(8), 62–80.

National Center for Education Statistics. (2011). Accessed through http://webcaspar.nsf.gov/, 2011.

Othman, M., & Latih, R. (2006). Women in computer science: No shortage here! *Communications of the ACM, 49*(6), 111–114.

Papastergiou, M. (2008). Are computer science and information technology still masculine fields? High school students' perceptions and career choices. *Computers & Education, 51*(2), 594–608.

Sax, L. J., Hagedorn, L. S., Arredondo, M., & Dicrisi, F. A. (2002). Faculty research productivity: exploring the role of gender and family-related factors. *Research in Higher Education, 43*(4), 423–446.

SJR. (2010). SCImago journal & country rank. http://www.scimagojr.com/journalrank.php.

Van Arensbergen, P., Van der Weijden, I., & Van den Besselaar, P. (2012). Gender differences in scientific productivity: A persisting phenomenon? *Scientometrics, 93*(3), 857–868.

Vela, B., Caceres, P., & Cavero, J. M. (2012). Participation of women in software engineering publications. *Scientometrics, 93*(3), 661–679.

Wainer, J., Eckmann, M., Goldenstein, S., & Anderson Rocha, A. (2013). How productivity and impact differ across computer science subareas. *Communications of the ACM, 56*(8), 67–73.

Xie, Y., & Shauman, K. A. (1998). Sex differences in research productivity: New evidence about an old puzzle. *American Sociological Review, 63*, 847–870.