Gender Differences in Public Code Contributions: a 50-year Perspective

Stefano Zacchiroli

Gender Differences in Public Code Contributions: a 50-year Perspective

Stefano Zacchiroli
Université de Paris and Inria, France

***Abstract*—Gender imbalance in information technology in general, and Free/Open Source Software specifically, is a well-known problem in the field. Still, little is known yet about the large-scale extent and long-term trends that underpin the phenomenon. We contribute to fill this gap by conducting a longitudinal study of the population of contributors to publicly available software source code. We analyze 1.6 billion commits corresponding to the development history of 120 million projects, contributed by 33 million distinct authors over a period of 50 years. We classify author names by gender and study their evolution over time.**
We show that, while the amount of commits by female authors remains low overall, there is evidence of a stable long-term increase in their proportion over all contributions, providing hope of a more gender-balanced future for collaborative software development.

GENDER IMBALANCE in science is an established and well-known phenomenon: women are underrepresented in STEM [3] and even more so in computing [6]. In the field of software development, Free/Open Source Software (FOSS) projects have been studied from the perspective of gender imbalance using various approaches.

Survey-based studies have repeatedly reported low women participation. Surveys up to 2003 [2] reported 95–99% man dominance in FOSS; a 2013 survey [10] observed a ratio of 10% women respondents. These surveys targeted FOSS contributors at large (with no restriction on project affiliation), relied on participant self-selection, and reached a maximum of several thousand usable responses each. Specific FOSS communities have also been studied for gender imbalance, e.g., Debian [7] or KDE [9], with similar results.

Quantitative studies of byproducts of collaborative FOSS development have analyzed selected projects to quantify gender imbalance in: mailing lists [5], support forums [14], GitHub teams [15] and pull requests [13]. They have confirmed the under-representation of female contributors in FOSS and also found evidence of measurable biases against them.

Paper contributions

A piece of knowledge that is still missing is a large-scale analysis of public code contributions, to establish a global breakdown of contributions by gender and to verify if long-term trends about gender participation exist in FOSS development. This paper contributes to fill these gaps. Specifically, we will address the following research questions:

RQ1. What is the overall breakdown by gender in contributions and contributors to public source code?

RQ2. Is there a long-term trend in the proportion of contributions and contributors to public source code by gender?

Answers to these questions will help confirming

Published by the IEEE Computer Society

or disputing past results on gender imbalance, this time at the unprecedented scale of public code. If stable trends were to be observed, they might also provide insights about what to expect in the future, informing policy making.

In order to answer the research questions we conduct a longitudinal study of the population of contributors to publicly available source code over a period of 50 years. To that end we retrieve the commits of more than 120 million collaborative projects from Software Heritage [1], totaling 1.6 billion commits. We then classify author names by gender using a frequency-based approach implemented on top of GENDER-GUESSER [12]. Finally we aggregate results by authors and number of commits, and analyze their evolution over time.

Replication package

A replication package for this paper is available from Zenodo at https://zenodo.org/record/4140789 (DOI: 10.5281/zenodo.4140789).

DATASET AND METHODOLOGY

We have retrieved from Software Heritage [1], [8] a snapshot of all the commits the project has archived until 2020-05-13. It consists of 1 661 391 281 commits (1.66 B), unique by SHA1 identifier, harvested from about 120 million public projects coming from major development forges (GitHub, GitLab, etc.) and source code distributions (Debian, PyPI, NPM, NixOS, etc.). For each commit we have its identifier, timestamp, and author full name.

We removed from the corpus commits with implausible timestamps, i.e., commits before the Unix epoch and commits "in the future" w.r.t. the date of the snapshot, with a tolerance of 1 day. Doing so excluded only 11 M commits (0.66% of the corpus). Figure 1 shows the number of commits in the corpus over time. It exhibits the already observed [11] exponential growth of public code (the notch for 2020 is a binning artifact due to the incompleteness of that year in the corpus).

The initial set of *distinct* authors associated to all commits consists of 33 660 524 (33.7 M) names. As most version control systems (VCS) do not store encoding information, author names in the dataset are raw *byte* sequences. We converted them to Unicode strings, trying the popular UTF-8 encoding and successfully converting 33 657 517 commits (99.991%).

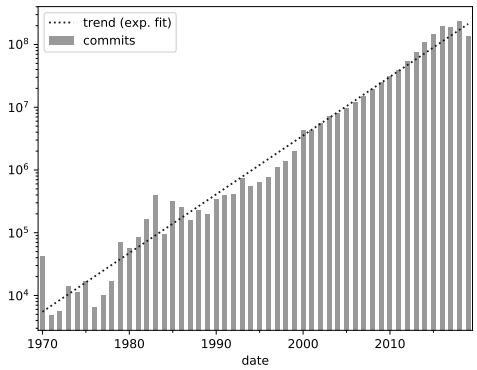


Figure 1: Total number of yearly commits by all authors. The log scale on the Y-axis highlights the exponential growth of public code.

We then filtered out implausible names such as: email addresses (used by mistake by authors *in lieu* of their name), names consisting only of blank characters, overlong names (more than 100 characters), and names containing more than 10% non-letter characters. This filtering reduced the corpus to 26 M authors after having removed: 7.5 M non-letter, 150 K emails, 25 K blank, and 31 overlong names. Finally we converted names to lowercase and normalized spaces, obtaining 13.2 M unique author strings.

Detecting the gender of a name is difficult in general [12] and even more so at this scale, geographic diversity, and lack of curation. Assigning a gender to a name also reinforces the gender binary, contributing to the marginalization of individuals who do not identify as men or women. A better approach is to ask authors for self-identification, but doing so is unfeasible at this scale. We hence delegate gender inference to automated tooling and we use the results only in aggregate form to study long-term trends. Throughout the paper we make no claim about gender identity (as in: the personal sense of one's own gender) and only discuss gender trends to the extent of which they can be inferred from author names.

Based on the results of a recent thorough benchmark of gender detection tools [12] we

have chosen GENDER-GUESSER, because it shines on heterogeneous inputs. GENDER-GUESSER is implemented in Python and is open source (https://pypi.org/project/gender-guesser/). This last point is particularly relevant: alternatives based on commercial APIs might give better accuracy, but would hinder replicability.

GENDER-GUESSER takes as input a Unicode string, which is supposed to be a *first* name, and returns the detected gender as one of 6 possible values, depending on the tool's certainty about the result: {*male*, *mostly male*, *unknown*, *mostly female*, *female*, *andy*} (the last one for unisex names).

Authors in our corpus are not split into first v. family name, but that distinction is not meaningful anyway in all the world cultures represented in the corpus [4]. Hence, to determine the gender of an author we apply a *majority criterion*. We use GENDER-GUESSER to determine the gender of each blank-separated *word* in the author name as a string. Then, if a strict majority of words are detected as belonging to one gender (no matter how strongly) we associate that gender to the entire author name; otherwise its gender will remain unknown, formally:

$$M_a = \{w \in a | guess(w) \in \{male, mostly\ male\}\}$$
$$F_a = \{w \in a | guess(w) \in \{female, mostly\ female\}\}$$

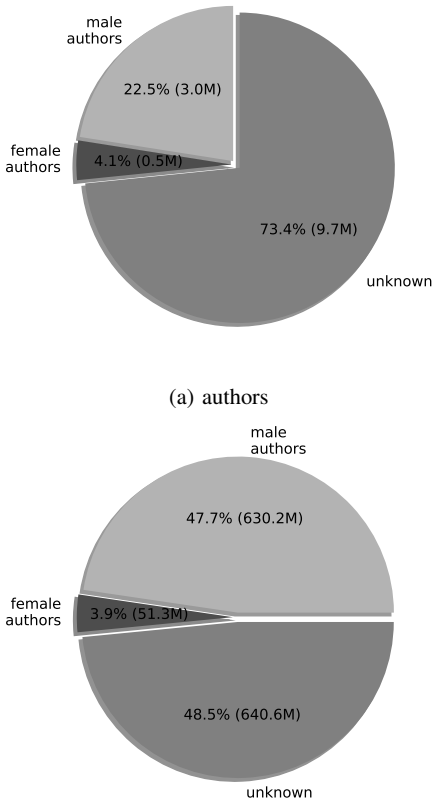
where a is an author name from our author corpus, w a word in that string, and $guess(w)$ denotes the invocation of GENDER-GUESSER. The gender of an author name a is then determined as follows:

$$GG(a) = \begin{cases} \male & if \quad |M_a| > |F_a| \\ \female & if \quad |F_a| > |M_a| \\ ? & otherwise \end{cases}$$

We can now partition the commit corpus $\mathcal{C}$ in the sets of commits by male authors, by female authors, or by authors for which we could not determine a gender, as follows:

$$\begin{aligned} \mathcal{C}_{\male} &= \{c \in \mathcal{C} \mid GG(c) = \male\} \\ \mathcal{C}_{\female} &= \{c \in \mathcal{C} \mid GG(c) = \female\} \\ \mathcal{C}_{?} &= \{c \in \mathcal{C} \mid GG(c) = ?\} \end{aligned}$$

We have computed these sets in practice, by running GENDER-GUESSER on each word in author names and determining the majority gender for each of them.



(a) authors

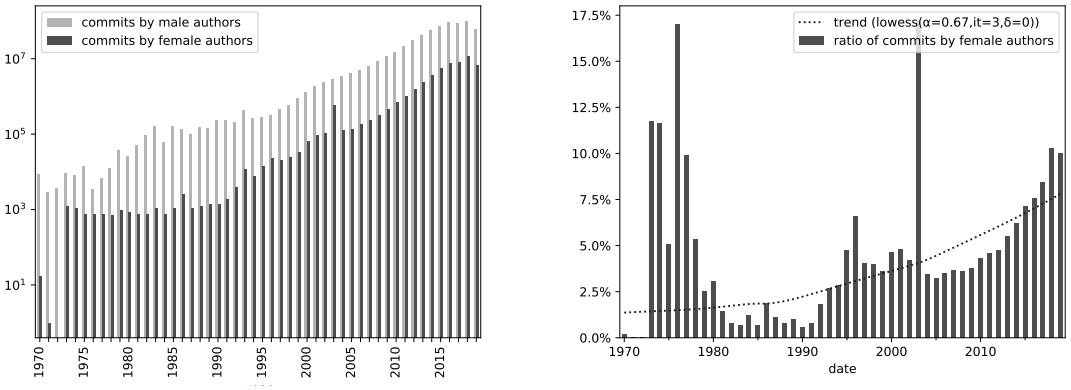
(b) commits

Figure 2: Breakdown of authors and authored commits by gender

RESULTS

Figure 2 shows the overall breakdown of detected genders in the studied corpus (RQ1). We were able to detect a gender for 3.5 M author names, or 26.6% of the author corpus. Author names with a detected gender account for 682 M commits, or 51.6% of the commit corpus. We have verified that the ratio of commits for which a gender could not be determined remains within 30–50% over time. Also, it has been shrinking for the past 20 years, during which the vast majority of commits have been produced (due to the exponential growth of the dataset).

Focusing on the author names for which we could determine a gender, 3 M (84.6%) are male authors v. 0.5 M (15.4%) female authors. In terms of contributions, commits by male authors are 630 M (92.5% of commits for which we could determine a gender) v. 51.3 M (7.5%) by female



(a) total number of yearly commits by author gender (note the log scale on the Y-axis)

(b) proportion of commits by female authors (on the total of commits for which gender could be determined)

Figure 3: Commits breakdown by detected gender over time

authors. In terms of diversity the picture is pretty dire: **male authors have contributed more than 92% of public code commits over the past 50 years**.

To answer RQ2, Figure 3a shows the evolution over time of commits authored by gender, excluding commits by authors for which we could not determine a gender. Consistently with the breakdown by gender in the corpus as a whole, we observe that the *yearly totals* of commits by female authors have lagged behind commits by male authors by significant margins for half a century.

However, female authors are increasingly contributing to public code. Figure 3b highlights this, showing the 50-year evolution of the ratio of commits by female authors over the total of commits for which we could determine a gender. The figure shows both yearly ratios as percentages and a locally weighted scatterplot smoothing moving regression over the entire period. **The ratio of commits by female authors has grown steadily over the past 50 years, reaching in 2019 for the first time 10% of all contributions to public code.**

Note also how the growth trend in the ratio of female-authored commits is steeper over the last 15 years (2005–2019) than before. This is significant because, due to the exponential growth of public code, those years have contributed the vast majority of commits to the entire corpus—and hence also contributed the most to the ongoing "catch up" in the total amount of commits by female authors v. commits by male authors.

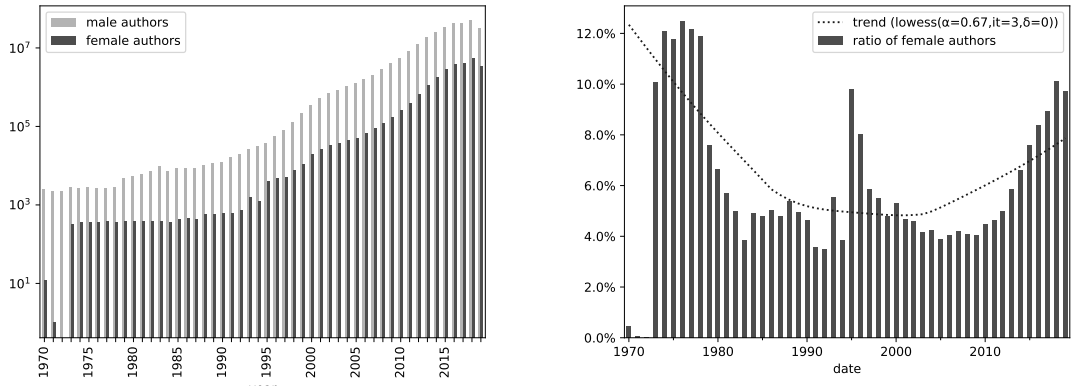
Figure 4 shows the yearly evolution of the number of *active authors* by gender, i.e., authors that have contributed at last one commit in a given year. In particular, Figure 4b confirms the **significant growth of active female authors from around 4% in 2005 to more than 10% of all public code authors in 2019**. If this trend is to continue, gender diversity among public code commits authors will increase significantly over the next few years.

DISCUSSION

To the best of our knowledge this is the first longitudinal gender study performed on public code at this scale—both in terms of population size and observed time period. The main trade-off in working at this scale is that we could not rely on curated gender information, e.g., author-provided information or interviews with them. Also, we had to work with non-parsed author names, leading to the need of using automated tools and heuristics.

Construct validity

The approach used for gender detection is crude. It is easy to come up with examples of family names composed by multiple words that are also common first names associated to a given gender, which will win majority over the gender detected for the *actual* first name. We do not expect this to happen often though. In general, family names are



(a) total number of yearly authors by gender (note the log scale on the Y-axis)

(b) proportion of female authors (on the total of authors for which gender could be determined)

Figure 4: Breakdown of authors with at least 1 yearly commit by detected gender over time

reported as unknown by GENDER-GUESSER, not contributing to shifting the detected author gender in either direction. We have verified this empirically during experiment design. Doing it more extensively would boil down to validating the accuracy of GENDER-GUESSER itself, which has already been done in the literature [12], supporting our tool choice.

A related threat is posed by usernames used instead of full names, which happen in old VCSs like CVS and Subversion. Previous considerations about family names apply to this case too. Also, 98.2% of the repositories in Software Heritage are from VCSs that support author full names (e.g., Git, Mercurial), so we do not expect this problem to be statistically significant.

Author names for which we could not determine a gender might impact our results. Given how unknown responses by GENDER-GUESSER do not affect gender assignment to authors, we consider this loss of coverage an acceptable consequence of our tool choice. As it is customary, we have excluded unknown gender authors from analysis. We could still determine author gender for almost 700 M commits, which remains an unprecedented scale for gender imbalance studies.

Qualitatively, the consistency of the observed trends with recent survey-based work [10] about the increase of women participation in FOSS further supports our results and vice-versa.

External validity

We do not claim to have analyzed the entire body of collaboratively developed software. We have nonetheless analyzed the largest publicly available corpus of commits coming from public version control system repositories. We do not think a much larger coverage is achievable without, for instance, adding large non-public forges (e.g., coming from large-scale inner source practices) to our sample, which would hinder replicability and impact on corpus diversity.

CONCLUSION

We have conducted the first large-scale longitudinal study of gender imbalance among authors of collaboratively developed, publicly available code. The study spans 1.6 billion commits harvested from 120 million projects and contributed by 33 million authors over a period of 50 years.

Results give a mixed message about gender diversity in public code collaboration. Overall, contributions by female authors remain scarce: less that 8% of commits for which we could detect a gender, confirming decades of gender imbalance in Free/Open Source Software (FOSS).

On the other hand, contributions by female authors appear to be on the rise and are rising faster than those by male authors. In 2019 and for the first time in half a century commits by female authors have reached 10% of yearly contributions to public code. Looking at active FOSS authors over time we find evidence of a similar sustained

growth and increasing speed. If the trend of the past 15 years is to continue, FOSS authors and their contributions might soon reach a level of gender diversity comparable to other fields.

The goal of this study was, on purpose, broad and longitudinal. As future work we intend to maintain the longitudinal angle, but drill down into specific software ecosystems to check if significant differences in gender participation trends exist and, if so, why. Our results also hint at other differences in participation by gender—e.g., commits per person over time and weekly participation patterns—which we also intend to explore in future work.

ACKNOWLEDGMENTS

The author would like to thank Jesus M. Gonzalez-Barahona for insightful discussions on this study, as well as Molly de Blanc and Antoine Pietri for comments on early versions of this paper.

REFERENCES

1. Jean-François Abramatic, Roberto Di Cosmo, and Stefano Zacchiroli. Building the universal archive of source code. *Communications of the ACM*, 61(10):29–31, September 2018.
2. Paul A David and Joseph S Shapiro. Community-based production of open-source software: What do we know about the developers who participate? *Information Economics and Policy*, 20(4):364–398, 2008.
3. Catherine Hill, Christianne Corbett, and Andresse St Rose. *Why so few? Women in science, technology, engineering, and mathematics*. ERIC, 2010.
4. Richard Ishida. Personal names around the world. https://www.w3.org/International/questions/qa-personal-names, 2011.
5. Victor Kuechler, Claire Gilbertson, and Carlos Jensen. Gender differences in early free and open source software joining process. In *8th International Conference on Open Source Systems, OSS 2012*, volume 378 of *IFIP Advances in Information and Communication Technology*, pages 78–93. Springer, 2012.
6. Jane Margolis and Allan Fisher. *Unlocking the clubhouse: Women in computing*. MIT press, 2002.
7. Mathieu O'Neil, Mahin Raissi, Molly de Blanc, and Stefano Zacchiroli. Preliminary report on the influence of capital in an ethical-modular project: Quantitative data from the 2016 debian survey. *Journal of Peer Production*, (10), 2017.
8. Antoine Pietri, Diomidis Spinellis, and Stefano Zacchiroli. The Software Heritage graph dataset: public software development under one roof. In *16th International Conference on Mining Software Repositories, MSR 2019*, pages 138–142, 2019.
9. Yixin Qiu, Katherine J. Stewart, and Kathryn M. Bartol. Joining and socialization in open source women's groups: An exploratory study of *KDE-Women*. In *6th International Conference on Open Source Systems, OSS 2010*, volume 319 of *IFIP Advances in Information and Communication Technology*, pages 239–251. Springer, 2010.
10. Gregorio Robles, Laura Arjona Reina, Jesús M. González-Barahona, and Santiago Dueñas Domínguez. Women in free/libre/open source software: The situation in the 2010s. In *12th International Conference on Open Source Systems, OSS 2016*, volume 472 of *IFIP Advances in Information and Communication Technology*, pages 163–173. Springer, 2016.
11. Guillaume Rousseau, Roberto Di Cosmo, and Stefano Zacchiroli. Software provenance tracking at the scale of public source code. *Empirical Software Engineering*, 25(4):2930–2959, 2020.
12. Lucía Santamaría and Helena Mihaljevic. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156, 2018.
13. Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, Chris Parnin, and Jon Stallings. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science*, 3:e111, 2017.
14. Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study. *Interacting with Computers*, 26(5):488–511, 2014.
15. Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. Gender and tenure diversity in GitHub teams. In *33rd annual ACM conference on human factors in computing systems, CHI'15*, pages 3789–3798, 2015.

Stefano Zacchiroli is Associate Professor of Computer Science at Université de Paris on leave at Inria, France. His research interests span formal methods, software preservation, and Free/Open Source Software engineering. He is co-founder and CTO of the Software Heritage project. Contact him at zack@irif.fr.