Gender Differences in Public Code Contributions

a 50-year Perspective

Stefano Zacchiroli
zack@irif.fr
@zacchiro

Université de Paris and Inria, France

26 August 2021
ESEC/FSE 2021
29th ACM Joint European SWE Conf. and Symposium on the Foundations of SWE

Context and Contributions

- Gender imbalance is a well-known negative phenomenon throughout STEM (science, tech., eng., math.)
- It is particularly severe in computing, and even more so in FOSS
  - large FOSS contributors surveys (2003–2017)
  - media analyses: mailing lists, forums, teams, PRs (2012–2017)

Contributions

1. Breakdown by gender in contributions to public source code (1.6 billion commits, 33 million authors)
2. Long-term trends (50 years) about the evolution of said contributions by gender

This is a journal first presentation of the following paper:

Stefano Zacchiroli

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

IEEE Software, 38(2):45-50, 2020

Dataset

- All commits archived by Software Heritage up to 2020-05-13
- Total: 1.6 billion public commits
- Cover 120 million public repositories from GitHub, GitLab.com and other major FOSS hosting platforms and distributions
- Commits include: timestamps and Git-style “fullnames”
  - e.g., `"Jane Doe <jane@example.com>"`

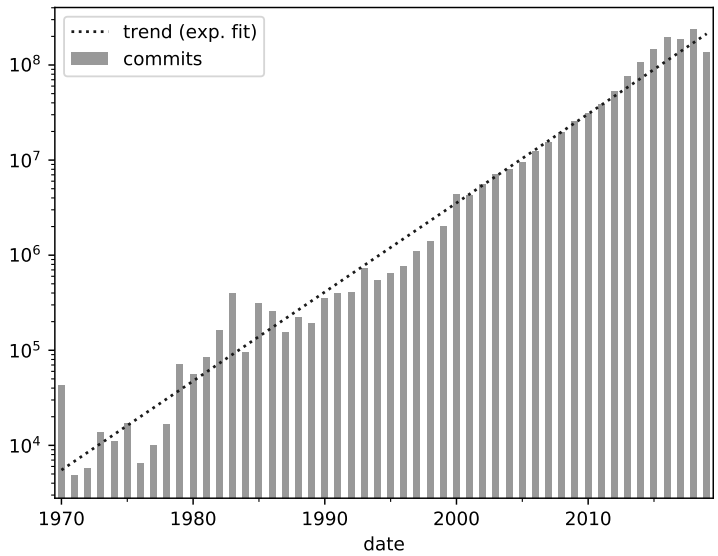

Figure: N. of commits by year, based on author timestamp (log Y-axis)

Methodology

- At this scale it is not feasible to interview contributors to *ask* what their gender is.
- We use `gender-guesser`, a frequency-based gender guesser for first names. It is widely used in the literature, open source, and has been shown to work well with international names.
- As Git full names do not separate *first* names, we apply the following majority criterion to detect author genders:
  1. Strip emails from full names
  2. Tokenize the rest splitting at blanks
  3. Detect the gender of *each token*
  4. Assign the majoritarian gender across tokens to the author

  Note: `gender-guesser` returns "unknown" when in doubt, which is the case for most non-first-name tokens.

Results (RQ1)

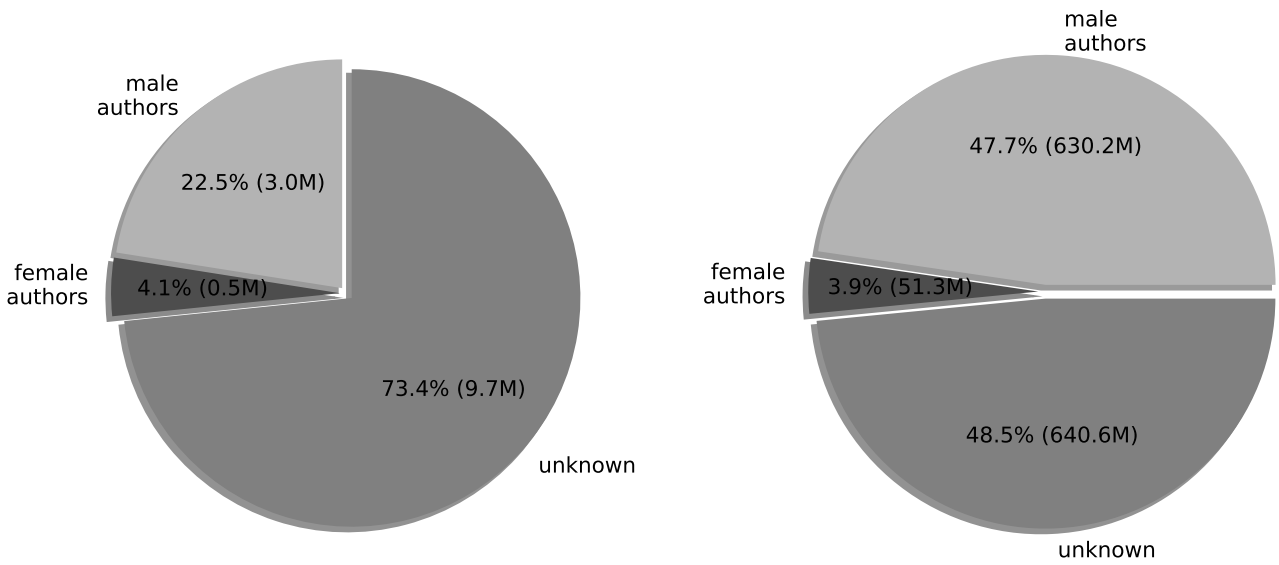

Figure: Breakdown of authors (left) and authored commits (right) by gender for the entire corpus. 85% of (detectable) authors of public code commits are male, 95% of (detectable) contributions are by male authors.

Key finding #1

Male authors have contributed more than 92% of public code commits over the past 50 years.

Results (RQ2)

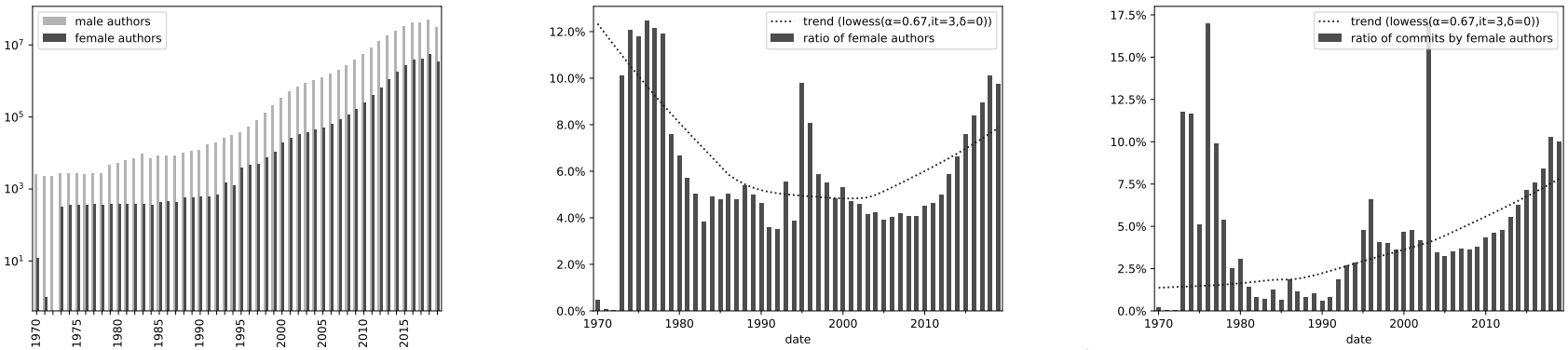

Figure: Breakdown of authors by detected gender over time: total commits (left) and proportion of female authors (middle). Counting by contributions confirms this long-term trend (right).

Key finding #2

- The ratio of commits by female authors has grown steadily over the past 20 years, reaching in 2019 for the first time 10% of all contributions to public code.
- Female authors have grown from 4% in 2005 to more than 10% of all public code authors in 2019.

- We have studied the evolution of the gender of authors of public code commits over 50 years at a very large scale (1.6 billion public commits).
- Bad news. We confirm the gender imbalance in FOSS: Male authors have contributed more than 92% of public code commits over the past 50 years.
- Good news. There is hope for a more gender-balanced future: The ratio of commits by female authors has grown steadily over the past 20 years, reaching in 2019 for the first time 10% of all contributions to public code.

Stefano Zacchiroli / zack@irif.fr / @zacchiro

These slides are available at:
https://upsilon.cc/~zack/talks/2021/2021-esecfse-gender.pdf
For further details check out the full paper:

Stefano Zacchiroli
Gender Differences in Public Code Contributions: a 50-year Perspective
IEEE Software, 38(2):45-50, 2020