Impact of Gender on OSS File Contributions

Leilani Torres
The College of Wooster
Wooster, OH, USA
ltorres25@wooster.edu

Heather Guarnera
The College of Wooster
Wooster, OH, USA
hguarnera@wooster.edu

Michael Collard
The University of Akron
Akron, OH, USA
collard@uakron.edu

Amber Garcia
The College of Wooster
Wooster, OH, USA
agarcia@wooster.edu

Abstract

We examine how gender impacts the use of specific programming languages, as analyzed across a stratified sample of 100k unique software developers from the World of Code (WoC) archive. A total of 50,000 male and 50,000 female developers are identified using the name-to-gender inference tool WikiGender-Sort. The top fifteen programming languages according to the 2024 StackOverflow Developer survey are considered. For each developer, we count the number of files that are edited in each programming language and compute the median across gender categories. Men and women tend to edit the same number of files among most programming languages, with the exception of developers using C#, C, Go, and Rust, which had more edits among men.

Keywords

—gender, diversity, open-source software, mining software repositories, software development, software ecosystems, World of Code

ACM Reference Format:
Leilani Torres, Heather Guarnera, Michael Collard, and Amber Garcia. 2025. Impact of Gender on OSS File Contributions. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2 (SIGCSE TS 2025), February 26-March 1, 2025, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3641555.3705198

1 Introduction

Women are underrepresented in STEM, particularly in computer science. Only 20% of software developers in the United States are women [1] and even less contribute to open-source software (OSS); less than 10% of new open-source accounts belong to women [6]. This under-representation in software development leads to environmental tech advancements tailored for a male demographic. It is crucial to discuss this gender gap to promote healthy environments for gender diversity.

Many previous studies have examined gender differences in different types of software contributions such as commits [7], lines of code, file edits, code reviews [4, 5], projects, and pull requests. This study seeks to challenge the stereotype of predominantly male contributions to OSS. We examine the top fifteen programming languages based on the 2024 Stack Overflow Developer Survey [2] to identify any gender disparities within contributions according to programming language. Specifically, we will address the following research question:

RQ1: How does gender influence the number of files edited within a given programming language? **RQ2**: What are the characteristics of languages that share gender similarities?

2 Methodology

We begin with a list of 51.4 million author names, aliases, and email addresses as identified from WoC servers, as of May 2024. We clean the dataset following the same strategy of Rossi and Zacchiroli [7]. Duplicates are removed based on author alias and email address pair (15.8 million), as well as blank names (0), names with over 100 characters (85), and names containing more than 10% non-letter characters (7.1 million).

Extensions	Language	Female	Male
js | js.map	JavaScript	8	9
html | htm	HTML	3	3
css | scss | less | sass	CSS	5	5
py | py3 | pyo |pyx | pyw | whl | pyd	Python	5	5
sql | sqllite | sqllite3 | mysql	SQL	2	2
ts | tsx	TypeScript	17	17
sh | zsh | bash | profile | bashrc | zshrc	Bash/Shell	2	2
java | iml | jar | dpj | xrb | aidl | mf | classpath	Java	9	9
cs | cspro	C#	17	19
c | h	C	6	7
php | twig	PHP	12	12
ps1	PowerShell	2	2
go	Go	4	5
rs | rlib | rst	Rust	4	5

Table 1: Programming languages with their file extensions and median number of files edited per gender category

After data cleaning, 28.4 million unique software developers remain. From this dataset, Wiki-Gendersort [3] inferred gender based on the developer's first name. Gender inference is difficult and risks marginalizing other identities such as non-binary individuals [8]. However, Wiki-Gendersort has shown [9] to be the most accurate (93.4%) free tool for gender categorization: M (masculine), F (feminine), UNI (unisex), INI (initials), and UNK (unknown). Out of 28.4

SIGCSE TS 2025, February 26-March 1, 2025, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0532-8/25/02
https://doi.org/10.1145/3641555.3705198

million authors, 46% were identified as male and 7% were identified as female. From this set, we obtained a stratified random sample of 50,000 male and 50,000 female developers.

We use the WoC API call *a2f* to return a list of all unique filenames edited by a developer through any of their commits. We use the extension of the filename, normalized to lowercase, to determine the programming language it was written in. We consider fifteen of the most popular [2] languages: JavaScript, HTML, CSS, Python, SQL, TypeScript, Bash/Shell, Java, C#, C, PHP, PowerShell, Go, and Rust. For each developer, a count of the number of files edited across each of the 15 languages is maintained. Table 1 illustrates the extensions considered for each programming language, and the median number of files edited by programmers in that language according to gender. To be considered a programmer in that language, a developer must have edited at least one file in that language. The number of programmers for each programming language according to gender is shown in Figure 2. Notably, programming language popularity differs from the 2024 Stack Overflow Survey. The difference may be due to analyzing open-source software from World of Code versus the developer platform from Stack Overflow (i.e., academia and industry).

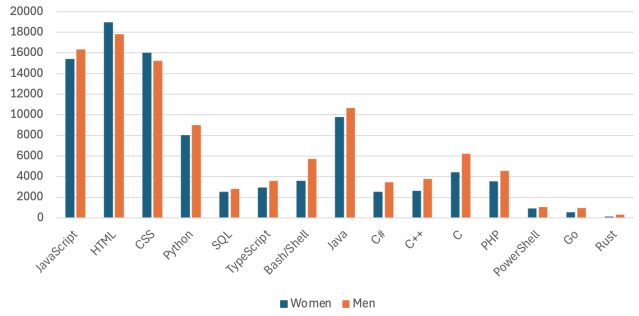

Figure 1: Number of programmers from our data set (50k male and 50k female) using the top 15 most popular languages, which are placed left-to-right in descending order of StackOverflow's ranking.

3 Discussion

Overall, the number of files edited among male and female developers is quite similar, suggesting equitable contribution. **To address our research question, we find that out of the top 15 programming languages, gender only influences the number of files edited for C#, C, Go, and Rust.** Programmers using TypeScript edit the most files (17), whereas Powershell programmers edit the fewest files (2). C# has the largest gender difference, with male developers editing 19 files while female developers edit 17.

Figure 1 demonstrates that the only languages with more female involvement than male involvement are HTML and CSS, which include 37.9% and 32.1% of all women, respectively; in comparison, there were 35.6% and 30.4% of all men who were HTML or CSS programmers, respectively. Our data suggests a different ranking of language popularity than that of the StackOverflow survey, which may better reflect the overall FLOSS community. TypeScript was originally ranked as the sixth most popular language, but our dataset demonstrates TypeScript as the ninth most popular language. It's possible that JavaScript would be ranked higher if other dialects (e.g. CoffeScript, Iced CoffeeScript, etc.) were also considered.

This partially dispels the stereotype of men contributing more than women. When analyzing male and female categories with the same sample sizes, there are not many differences. Regardless of the differences in popularity ranking, male and female developers tend to have similar trends in the usage of programming languages. These results suggest that there are only few differences between male and female participation when using equal representation (i.e., using a stratified data set which consists of half men and half women, rather than a random sample of the FLOSS population which is heavily male-dominated).

This study has a few limitations and other considerations. Although Wiki-Gendersort is highly accurate, there are challenges with non-English alphabet names [9] and it is applied to usernames that may not accurately define one's gender. Self-reported gender would be most accurate, but that data is not available. We also emphasize that our study does not measure the quality of contributions, only the quantity.

Future work may address some of these challenges through alternate gender inference tools with greater accuracy and greater analysis of the content of the contributions, such as evaluating file content. In addition, there is potential future work that considers the number of files edited in more depth. For example, comparing the number of developers exposed to a particular language (edited at least one file) and experienced with a given language (edited a number of files). We also plan to evaluate developer contributions with a greater level of granularity, including the number of commits and number of projects.

4 Acknowledgements

This work was partially supported by the Sophomore Research Program at The College of Wooster.

References

[1] [n. d.]. https://bls.gov/cps/cpsaat11.htm "Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity," Bureau of Labor Statistics. [Online]..

[2] [n. d.]. https://survey.stackoverflow.co/2024/ "2024 Developer Survey," Stack Overflow. [Online]..

[3] N Berube, M Sainte-Marie, and V Lariviere. 2020. Wiki-Gendersort: Automated gender detection using first names in wikipedia. *OSF* (2020). https://doi.org/10.31235/osf.io/ezw7p

[4] E Murphy-Hill and et al. 2023. Systemic gender inequities in who reviews code. *Proc. ACM Hum.-Comput. Interact* (2023). https://doi.org/10.1145/3579527

[5] R Paul, A Bosu, and K Sultana. 2019. Expressions of sentiments during code reviews: Male vs female. *IEEE 26th International Conference on SOftware Analysis, Evolution, and Rengineering(SANER)* (2019). https://doi.org/10.1109/SANER.2019.8667987

[6] Gede Prana, Denae Ford, Ayushi Rastogi, and et al. 2021. Including everyone, everywhere: Understanding opportunities and challenges of geographic gender-inclusion in OSS. *IEE Transactions on Software Engineering* (2021). https://doiAmiangshuBosu.org/10.48550/arXiv.2010.00822

[7] D Rossi and Zacchiroli. 2022. Worldwide gender differences in public code contributions and how they have been affected by COVID-19. *IEEE/ACM 44th International Conference on Software Engineering* (2022). https://doi.org/10.1145/3510458.3513011

[8] L Santamaria and H Mihaljevie. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput Sci* (2018). https://doi.org/10.7717/peerj-cs.156

[9] P Sebo. 2021. Performance of gender detection tools: A comparative study of name-to-gender inference services. *J Med Libr Assoc* (2021). https://doi.org/10.5195/jmla.2021.1185