Gender Differences in the Development of R Packages on GitHub

Carol Moore
USG
cm6ag@virginia.edu

Uyen Nguyen
University of Virginia
gmd8sq@virginia.edu

Gizem Korkmaz
Westat
gizemkorkmaz@westat.com

1 Introduction

The analysis of the gender dynamics in scientific research and respective outputs is crucial for ensuring that science policy is inclusive and equitable. Similar to other research outputs such as publications and patents, open source software (OSS) projects are also developed by contributors from universities, government research institutions, and nonprofits, in addition to businesses. Despite its reach and continued rapid growth, reliable and comprehensive survey data on OSS does not exist, limiting insights into contributions by gender and policymakers' ability to assess trends in gender representation.

Like in scientific research, the inclusion of diverse perspectives in software development enhances creativity and problem-solving. Using GitHub data, researchers have found positive correlations between gender diversity of an OSS development team and its productivity (Vasilescu et al., 2015; Ortu et al., 2017). Yet there is evidence of gender bias, with women facing higher standards to have their contributions accepted (Terrell et al., 2017; Imtiaz et al., 2019).

This exploratory study aims to quantify gender differences in development and use (impact) of OSS using publicly available information collected from GitHub. We focus on software packages developed for programming language R, with the majority of contributors from academia. The paper asks (1) what are gender differences in the volume of contributions? (2) has gender representation shifted over time? (3) is there a correlation between the gender of contributors and the impact of a package?

2 Data and Methods

We collected data on R installable libraries/packages from the Comprehensive R Archive Network (CRAN). The registry includes 19,700 packages and associated metadata such as contributor names, license, technical dependencies, and source code URL. We collected repository-level information (coding activity, contributor profiles, and repository attributes) for the packages that are developed and maintained on GitHub.

The dataset includes 1,883,977 commits to 7,016 registered R packages from 2008 to mid 2023 and information about 14,311 unique contributors. We also collected download counts for the respective packages.

2.1 Gender Classification

We used gender-guesser, an open-source Python package, to assign a gender label to contributors based on first name, ultimately categorizing them into male, female, non-binary or unknown genders. We classified 6,692 contributors as male, 755 as female and the remainder as non-binary or unknown.

2.1.1 Validation

The accuracy of gender-guesser is competitive with that of other name-based gender classification tools (Santamaría and Mihaljević (2018)), but such tools carry risk of error, especially for non-Western names (Santamaría and Mihaljević (2018), Sebo (2022)). GitHub has recently created a profile field for voluntary sharing of pronouns, which we used to evaluate the tool's performance.

Table 1 compares the gender-guesser prediction to self-reported pronouns. Accuracy is 97% when non-classification is excluded and 60% when included.[1] Precision for females was 89% and for males 98%.

		Predicted			
		female	male	unknown	total
Actual	female	33	4	36	73
	male	4	244	144	392
	total	37	248	180	465

Table 1: Validation of gender-guesser's classification using self-reported pronouns on GitHub.

Although these results are promising, nearly one-half of the users were not classified. Conclusions rest on an assumption that the actual gender mix of "unknown" users is close to the mix of identified males and females.

2.2 Dependency Network

While most packages in the CRAN registry only require base R to perform, some depend on other packages. A package's network centrality measures the extent to which other packages depend on it, suggestive of both its technical utility and economic impact. A directed edge $i \rightarrow j$ indicates that the package i depends on j to function. We obtain a network with 5,894 nodes and 9,128 directed edges. On average, the number of edges (incoming or outgoing) per node in the dependency network is 3.1. About 1,000 nodes do not point to (are not dependent on) another node. About 4,500 do not have any incoming links (nothing depends on them). This is consistent with our initial view that R packages tend to be specialized for specific scientific fields and methods.

We examine a number of centrality metrics to capture the impact of packages. Here, due to space limitations, we focus on indegree (i.e., the number of reuses a package has). We find that the package with the highest indegree centrality is ggplot2, with a score (incoming links / (n - 1)) of 7.9%. This is followed by MASS, Matrix, survival, and DBI. Cumulatively,

[1] Non-classification was mainly due to users from Asian locations, unisex western names, and organization accounts.

these five packages account for a quarter of dependencies in the R packages universe.

To analyze the correlation between network impact and GitHub contributions by gender, we extract centrality metrics for the packages that were developed in GitHub. Over 99 percent of the approximately 7,000 GitHub-developed R packages appear in the CRAN listing. Among these, about 1,500 are in the R dependency network. Other GitHub-developed R packages are singletons to which we assigned network centrality scores of 0. Excluding singletons, the mean node had 2.65 incoming links. As in the broader R network, only a few packages had non-negligible centrality scores.

2.3 Other Metrics of OSS Contribution and Impact

In addition to network centrality metrics, we use downloads from CRAN (indicative of popularity among R users) and the number of forks and stargazers on the package's repository (developer interest) to capture impact. Downloads, forks and stargazers were reported as of mid-2023 and are normalized for package age. We used package ownership (suggestive of leadership), and the number of commits and lines of code added per month of tenure.

3 Results

Through percentage breakdowns we showcased how different gender groups contributed to OSS projects through commits, lines of code, and package ownership. To examine the link between gender and package impact, we correlated impact metrics with the percentage of a package's code additions that were committed by female developers and compared these metrics by gender of the package's owner. Finally, we documented growth rates and counts of contributors by gender since GitHub's start in 2008 to determine if the gender gap has narrowed. Our results are summarized below.

Figure 1 illustrates growth in package ownership by gender over time. Although female representation has grown, male developers continue to dominate ownership.

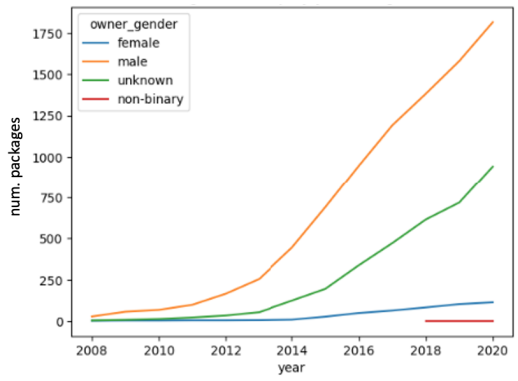

Figure 1: Number of packages owned by male and female developers by year

Figure 2 compares the distribution of impact metrics of packages owned by male and female developers. We found no significant difference in respective impacts.

Overall, we found that metrics were highly skewed (follows a power-law distribution), with a few observations having an out-sized impact on the R ecosystem. The topmost contributors in terms of volume (number of lines of code, number

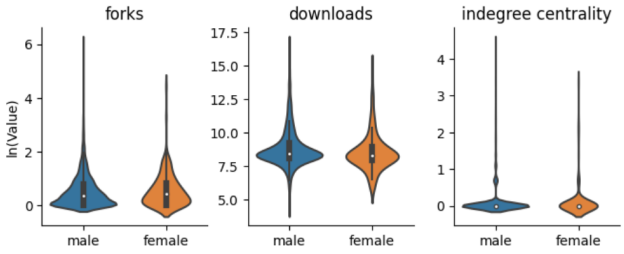

Figure 2: Distributions of package impact metrics for packages owned by make vs female

of packages contributed to) are male. However, the median female has the same or greater volume of commits, packages, and lines of code added per month as the median male.

In 2008, only 5% of contributors were female. That figure crept up to 11% in 2020. Moreover, teams have grown more diverse: the average package team was 1%-2% female during 2008-2014 and 4%-5% 2016-20. Nevertheless, the absolute difference in the number of male and female developers has grown. We found no evidence that a package's impact is correlated to the percentage of lines of code added by females, or that it differs by the gender of the package owner.

4 Discussion

Gender classification and analysis shed light on gender representation and contributions in the rapidly growing and valuable OSS sector, providing insights for further research and decision-making. This exploratory study has limitations due to missing gender information for a significant number (about a half) of developers. We recommend future research prioritize gender classification to reduce uncertainty in male-female comparisons and trend analysis. Options include named entity recognition of the national origins of names to better label the dataset and identify candidates to be discarded, and recurrent neural networks (e.g., (Hu et al., 2021)).

5 Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant Number 2306160. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Yifan Hu, Changwei Hu, Thanh Tran, et al. 2021. What's in a name?–Gender classification of names with character based machine learning models. *Data Mining and Knowledge Discovery*, 35(4):1537–1563.

Nasif Imtiaz, Justin Middleton, Joymallya Chakraborty, et al. 2019. Investigating the effects of gender bias on GitHub. In *2019 IEEE/ACM 41st ICSE Conference*, pages 700–711.

Marco Ortu, Giuseppe Destefanis, Steve Counsell, et al. 2017. How diverse is your team? investigating gender and nationality diversity in GitHub teams. *Journal of Software Engineering Research and Development*, 5(1):1–18.

Josh Terrell, Andrew Kofink, Justin Middleton, et al. 2017. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science*, 3:e111.

Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, et al. 2015. Gender and tenure diversity in GitHub teams. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*.