CHAPTER 14

How Much Do Women Build Open Source Infrastructure?

Huilian Sophie Qiu, Northwestern University, USA.*

Zihe H Zhao, Rice University, USA.

Tielin Katy Yu, Carnegie Mellon University, USA.

Laura Dabbish, Carnegie Mellon University, USA.

Bogdan Vasilescu, Carnegie Mellon University, USA.

There have been many past reports of women being underrepresented among contributors to open source, from surveys and analyses of repository data. In this chapter we take a fresh, comprehensive look at the representation of women in open source, focusing on historical trends among *infrastructure* projects – the libraries and packages indexed by popular package managers that so much of the world relies on. We start by compiling and synthesizing existing empirical data from the literature and then use an automatic name-based gender inference technique to capture population level across 20 open source package manager ecosystems. Our results reveal a promising upward trend in the percentage of women among both highly active ("core") and general repository contributors over time, but also high variation in the percentage of women contributors across ecosystems. The chapter is based on a short paper we presented at ICSE SEIS 2023 [44].

Equity, Diversity, and Inclusion in Software Engineering, https://doi.org/10.1007/978-1-4842-9651-6_14

Introduction

The economic value and importance of open source software (OSS) to the economy and society as a whole are, by now, well recognized. Companies big and small, nonprofits, government entities, scientists, students, and hobbyists all use OSS libraries and packages [18]. To maintain all this digital infrastructure, a constant supply of effort is needed, often by volunteers, to fix bugs, patch vulnerabilities, and implement new features. Prior research has repeatedly shown that the availability of this effort should not be taken for granted – open source contributors can choose to disengage at any time for a variety of reasons [37], and even widely used, popular projects can end up being maintained by no one at all [4, 12].

Among the challenges to open source software sustainability, low gender diversity is particularly problematic because it hinders the benefits that a team could have possessed otherwise. It is beyond being a problem of social justice, as there is plenty of evidence demonstrating the benefits of having a gender-diverse team. For example, evidence shows that having a gender-diverse team in public code collaboration could enhance productivity and lower community smells [11, 46]. One reason behind the better performance is that men and women tend to display different personalities [49]. Leveraging positive personality traits that are associated with better team performance can lead to more successful teams [59]. At the same time, a diverse team can better understand the needs of their users, which are often diverse [38].

Practitioners and researchers have been working on solving the problem of low gender diversity in OSS. Many studies and reports in the past two decades showed low representation of women in OSS (see the section "Related Work" for a review). Active research areas include identifying roles women play in OSS development [54], detecting barriers that women face when entering OSS [17], and quantifying biases women face when making contributions [53]. In practice, there are initiatives to remove barriers for women and to create more inclusive communities, such as Open Source Diversity,[1] Outreachy,[2] and Rails Girls Summer of Code.[3]

There have been many attempts to assess the gender representation in the open source software community. Although prior studies reached a general agreement on the overall low fraction of women in the population, the reported percentages have a

[1] https://opensourcediversity.org
[2] https://www.outreachy.org
[3] https://railsgirlssummerofcode.org

high variance, the possible causes of which could be unrepresentative samples, different subpopulations, different time periods, or different methods. In this chapter, we add one large-scale study to the literature while fixing the method and looking over time and across ecosystems. This chapter is a descriptive study that reports the representation of women in OSS slicing by three dimensions. We first slice data over time to show how the gender distribution evolves. Then we slice data by ecosystem, since each of them has different management practices [6]. We also segment the population vertically to analyze women's distribution among core contributors, those who are more experienced and responsible for the majority of the contributions [28].

When investigating gender distribution, we followed many previous studies [48, 57] and used automated gender inference tools to infer genders based on the information disclosed by contributors, oftentimes names. These methods have certain known limitations and biases, including the imperfect accuracy and the assumption of *binary* gender, which does not reflect the current perception of gender [50]. We are aware that the use of the inference on individuals can be harmful [26, 29]. Therefore, our study only uses name-based gender inference on the population level and treats the results as only an approximation of the real situation [35].

Related Work

Automatic Gender Inference Tools

Researchers have explored various techniques to automatically infer gender of individuals. This section discusses the approaches available to our GitHub source data. Note that all classifiers here assume binary gender, and their benchmarks also consist of only data of binary gender.

Appearance-based gender inference has been extensively studied in the field of computer vision, where many classifiers can achieve an accuracy higher than 90%, even 99% [2] or nearly 100% [62]. However, a large number of GitHub users are using default profile pictures, and there is no guarantee that a contributor's profile picture is a picture of themselves. Hence, we did not use appearance-based inference because the results would be very unreliable.

Researchers have also explored *text-based gender inference*, which relies on vocabulary and frequency of words [34] and even style markers and structural characteristics [13]. However, our text pieces on GitHub, such as commit messages, are usually short, and the accuracy of this technique is low.

To the best of our knowledge, *name-based gender inference* is the most commonly used approach in the software research community. Certain tools perform the inference based on only an individual's first name. For example, Gender-guesser is a Python package that uses the first name to assign "unknown," "andy" (androgynous), "male," "female," "mostly_male," or "mostly_female" to an individual. In comparison, several tools incorporate one's geolocation or cultural origin into their inference. For example, both Namsor and NameAPI are paid services that infer one's cultural origin based on their last name. Based on benchmark evaluations by Santamaría and Mihaljević [50] and Sebo [51], Gender API and Namsor are the most accurate tools with accuracy higher than 90%. Thus, we pick Namsor as our gender inference tool.

Researchers have started reflecting on the negative impact of automatic binary gender inference tools. Hamidi et al. [26] criticized the tools' assumption of binary gender as "gender reductionism." We acknowledge and agree that the limitation also exists in name-based gender inference, including ours, and caution against using such technology to make individual-level inferences. As we argued previously, we only make population-level inferences to get a general sense of global trends and differences among ecosystems.

Gender Distribution from Prior Studies

With rising awareness of the low gender diversity problem, many studies have attempted to estimate the gender composition in the OSS community. Although all studies report a low percentage of women contributors, these numbers have wide variation ranging from 1% to 12%. Building on the overview or women ratios across years by Trinkenreich [55], we provide an overview of the results reported by prior studies grouped by methods.

Surveys: The first section of Table 14-1 lists the studies that rely on survey data to measure gender distribution. Surveys can capture people's self-identified gender and arguably increase the precision of gender identification [36]. However, survey data, albeit more reliable and accurate, are prone to selection bias [5]. Moreover, survey samples are usually small, making it hard to generalize.

Table 14-1. *Women ratios in prior works grouped by data sources and methods*

Year	Source	Sample Size	Ratio	Citation	Project
		Gender Ratios Reported from Survey Data			
2001	Online survey	5,478	0%	Robles et al. [47]	
2002	Online survey	2,784	1.1%	Ghosh et al. [24]	
2001–2002	Email	684	2.5%	Lakhani et al. [32]	
2002	Email	79	5%	Hars and Ou [1]	
2003	Online survey	1,588	1.6%	David et al. [15]	
2013	Online survey	2,183	10.35%	Robles et al. [46]	
2015	Online survey	816	24%	Vasilescu et al. [59]	
2017	Online survey	6,000	5%	GitHub [23]	
2017	Online survey	64,000	7.6%	Stack Overflow [52]	
2019	Online survey	119	10.9%	Lee et al. [33]	
2021	Online survey	242	7.6%	Gerosa et al. [22]	
		Gender Ratios Reported from Mining Software Repositories			
2012	Email subscribers, US census	1,931	8.27%	Kuechler et al. [31]	
2012	Stack Overflow	2,588	11.24%	Vasilescu et al. [57]	
2015	GitHub, genderComputer [57]	1,049,345	8.71%	Kofink [30]	
2015	GitHub, genderComputer	873,392	9%	Vasilescu et al. [60]	
2017	GitHub, social media	328,988	6.36%	Terrell et al. [53]	

(*continued*)

Table 14-1. (*continued*)

Year	Source	Sample Size	Ratio	Citation	Project
2017	OpenStack, genderize.io[4]	-	10.4%	Izquierdo et al. [27]	
2019	GitHub, Namsor [9]	300,000	9.7%	Qiu et al. [42]	
2019	Gerrit, genderComputer, social media	4,543	8.8%	Bosu and Sultana [7]	
2020	GitHub, genderComputer, Namsor	1,954 core	5.35%	Canedo et al. [8]	
2021	GitHub, genderComputer, Simple Gender [21]	1,634,373	5.49%	Vasarhelyi et al. [56]	
2021	GitHub, genderize.io	65,132	10%	Prana et al. [41]	
2022	Software Heritage [40], gender-guesser[5]	21.4M	10%	Rossi et al. [48]	
	Gender Ratios Reported from Different Ecosystems or Projects				
2014	Mailing list	3,342	9.81%	Vasilescu et al. [58]	Drupal
2014	Mailing list	3,611	7.81%	Vasilescu et al. [58]	WordPress
2016	Online survey	765	5.2%	Sharan [20]	Apache
2005–2016	GitHub	14,905	8%	Cortázar [14]	Linux
2016	Online survey	1,479	2%	Raissi et al. [45]	Debian
2019	GitHub, Namsor	1,601	3.4%	Asri and Kerzazi [3]	Angular.js
2019	GitHub, Namsor	1,824	3.5%	Asri and Kerzazi [3]	Moby

(*continued*)

[4] www.genderize.io

[5] https://pypi.org/project/gender-guesser/

Table 14-1. (*continued*)

Year	Source	Sample Size	Ratio	Citation	Project
2019	GitHub, Namsor	3,723	4.2%	Asri and Kerzazi [3]	Rails
2019	GitHub, Namsor	1,672	5.3%	Asri and Kerzazi [3]	Django
2019	GitHub, Namsor	1,127	4.2%	Asri and Kerzazi [3]	Elasticsearch
2019	GitHub, Namsor	1,735	5.8%	Asri and Kerzazi [3]	TensorFlow
2019	Gerrit, genderComputer	258 core	3.87%	Bosu and Sultana [7]	Android
2019	Gerrit, genderComputer	151 core	3.97%	Bosu and Sultana [7]	Chromium OS
2019	Gerrit, genderComputer	24 core	4.17%	Bosu and Sultana [7]	Couchbase
2019	Gerrit, genderComputer	90 core	7.77%	Bosu and Sultana [7]	Go
2019	Gerrit, genderComputer	68 core	1.47%	Bosu and Sultana [7]	LibreOffice
2019	Gerrit, genderComputer	60 core	10%	Bosu and Sultana [7]	OmapZoom
2019	Gerrit, genderComputer	34 core	2.94%	Bosu and Sultana [7]	oVirt
2019	Gerrit, genderComputer	159 core	3.12%	Bosu and Sultana [7]	Qt
2019	Gerrit, genderComputer	73 core	4.1%	Bosu and Sultana [7]	TYPO3
2019	Gerrit, genderComputer	19 core	0%	Bosu and Sultana [7]	Whamcloud
2021	Online survey	2,350	14%	Carter et al. [10]	Linux

Mining software repositories: The second section of Table 14-1 lists the studies that rely on data mining to report gender distribution. In these quantitative studies, researchers often need to infer gender because not all platforms collect users' gender and not all users disclose their genders online. Thus, automatic gender inference tools have become a common practice. Despite the limitations, gender inference based on mined user information provides a more representative, larger-scale sample than the survey approach. It also eliminates the burden on the survey respondents and the efforts taken to collect survey results.

Ecosystems: The last section of Table 14-1 lists studies that report gender ratios in specific software ecosystems. The percentages of women range from 0% (Whamcloud) to 10% (OmapZoom) [7]. However, to the best of our knowledge, there is not a study that covers all major ecosystems, and many of the previous studies focus on a selection of projects rather than the entire ecosystem.

Methods

To conduct an ecosystem-level census, we used data from GHTorrent and retrieved the list of projects in the 20 largest package managers on libraries.io,[6] a service collecting data of open source packages. We only selected the 20 biggest package managers out of the total 38. Because our automatic gender inference is not perfect and can be used only as a population-level approximation, results in smaller ecosystems can fluctuate and become unreliable. We used data from GHTorrent [25], which provides trace data from GitHub between January 2008 and March 2021. However, we note the limitation that the data between June and December 2019 are missing.

Data Processing Pipeline

Extracting the list of open source infrastructural projects: We consider a GitHub project that is registered at libraries.io as an OSS project. Using the January 12, 2020, version of the dataset from libraries.io, which consists of entries of open source projects registered by the date, we parsed out 1,550,273 unique, valid projects that can be found on GHTorrent.

[6] https://libraries.io

Collecting contributions: Due to data traceability, we consider only commits, both code and documentation, as contributions. We acknowledge that this simplification neglects contributions such as management, avocation, and mentorship [54, 55]. However, many of these non-code activities are either untraceable or hard to quantify. Therefore, at this moment, we focus on only tractable contributions.

De-aliasing user entries: Because developers sometimes use different accounts when authoring commits in a project, we perform identity merging through a set of heuristic rules to ensure that we do not over-count users. Our de-aliasing method relies on user-level information, for example, emails and names [19, 61].[7] For example, if two accounts use the same email and similar names, that is, some or all parts are the same but in different orders, or the same name with similar emails, that is, their emails contain part of their names, their commits could most possibly be credited to one author.

Removing bots: To reduce the impact of bot contribution, we manually evaluate the activity of all users who made at least 1,000 commits in each ecosystem [16]. We found 511 unique bot accounts, which made 5,828,940 commits in total.

Aggregation granularity: To study how women's participation changes over time, we aggregate data into *three-month windows*, which ensures sufficient interactions among contributors since activities on GitHub are more sparse than those in companies. For windows that have less than 30 contributors whose genders can be inferred, we consider those windows as no activity, as the percentage of women might surge and become an outlier in the data.

Identifying core contributors: Adapting from the validated count-based methods by Joblin et al. [28] and Bosu et al. [7], we identified core contributors in the following way. For each ecosystem, within each three-month window, we first identified projects whose number of commits ranked top 10% in the ecosystem. Then, within each of the top projects, we identified each project's core developers as those who made more than 10% of the commits within that three-month window. In summary, in our analysis, a core contributor makes more than 10% of the commits in a project whose number of commits ranks top 10% in that ecosystem. We are specifically interested in core developers because cores typically take more responsibility for public project code contribution.

[7] https://github.com/bvasiles/ght_unmasking_aliases

Gender Inference

Of the 45,838,860 GitHub users in GHTorrent, 53.65% do not provide a name, and 3.84% are organizational accounts. We label these users' gender as *Unknown*. We also label users whose names have more than four parts (71,367 (0.16%)) as *Unknown* since a manual checking showed that most of them are names of organizations. We preprocess the remaining users' names by removing punctuations, common titles or prefixes, emails, and URLs.

Then, we infer the gender of each user with Namsor [9], one of the name-based gender inference tools with the highest accuracy [50, 51]. The tool makes inferences based on the first name and the cultural origin of the last name.

Namsor also provides a confidence level that a user's gender is correctly identified. We denote users whose gender inference confidence is lower than 0.7 as *Unknown* gender. Removing inferences with low confidence can increase the overall accuracy of our gender classification, yet setting a high confidence threshold cuts down our data size. Thus, we choose 0.7 as the threshold to retain 83.81% of the gender data. Of 1,823,414 users who have contributed to OSS projects, 911,990 (50.02%) are labeled as men and 54,859 (3.01%) as women. To reduce the effect of *Unknown* gender on our result, we calculate women fraction by

$$\frac{Number\ of\ Women Contributors}{Number\ of\ Women + Men Contributors}$$

Results

Gender Distributions in OSS and Different Ecosystems

Figure 14-1 shows the overall gender distribution in OSS libraries and its evolution over time. Overall, the percentage of women has been constantly low – no higher than 5.0%. Moreover, the percentage of women among all contributors in OSS projects is lower than that among core contributors.

For the gender distributions in the top 20 most popular OSS ecosystems and their evolution, we observed different patterns in different ecosystems. Due to the space limit, we display only plots from four more representative ecosystems in Figure 14-2: npm, CRAN, PlatformIO, and CPAN.

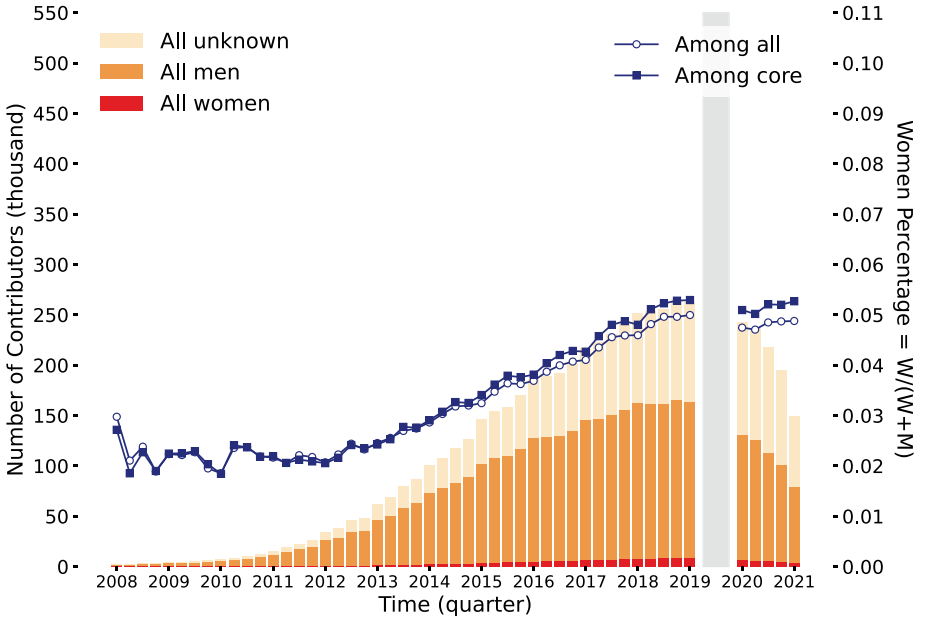

Figure 14-1. *Gender representation in OSS contribution overall. The gray bar covers the period where GHTorrent has missing data.*

For more figures, please visit our GitHub page.[8]

Figure 14-2a shows the trend of women percentage in the npm ecosystem. The pattern of npm's women percentage change is representative of many ecosystems, such as PyPI, Bower, and Go. Although the overall women percentage has been low all the time (lower than 6%), there is a steady increase overtime.

While most ecosystems exhibit increasing women percentage, the numbers are all lower than 10%, with the exception of CRAN, which reached 10.02% in 2021 (Figure 14-2b). CRAN is the package manager for the R programming language, which is widely used among academic researchers. The higher women percentage in CRAN may be due to the fact that the population of R users is more diverse because they come from various disciplines other than computer science [6].

[8] https://github.com/CMUSTRUDEL/OSS-gender-census-SEIS2023

Moreover, as shown in Figure 14-2c and 14-2d, PlatformIO and CPAN display a puzzling periodicity and minimum growth over the years. This pattern can be due to the fact that PlatformIO is a smaller ecosystem in our dataset. As a result, a small change in team composition can result in a large fluctuation. This also explains why we chose to only present results for the 20 larger ecosystems: the smaller the ecosystem is, the more likely it would be influenced by small changes.

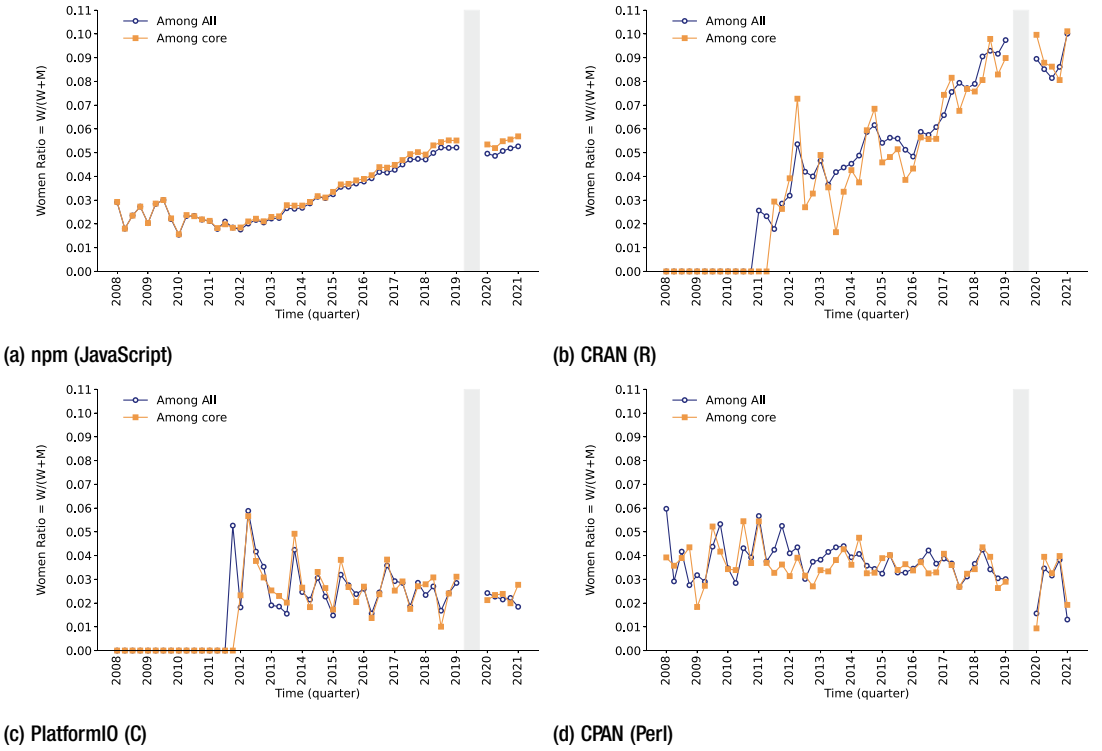

Figure 14-2. *Women distributions overall and in selected ecosystems. Gray bars cover the period with missing data on GHTorrent.*

For most ecosystems, the percentage of women exhibited an uphill pattern and reached its peak between 2018 and 2021. However, some languages commonly used for system programming – Perl, Rust, and C++ – reached their maximum percentage before 2014. Table 14-2 shows the percentages of women at the end of our data (January–March 2021) and the window during which the maximum percentage of women contributors occurred.

Gender Distributions Among Core Contributors

Starting with women percentage of 2.13% among core contributors and 2.25% among all contributors, the number has been steadily growing between 2008 and 2021. We observed that, while the women percentage among all contributors was higher than among cores in 2008, the difference between them was less than 0.01% in 2014. Between 2014 and 2021, we found that the women percentage among cores has surpassed that among all, leaving a slight but approximately stable margin of 0.3%.

Lastly, comparing the percentage of women among core contributors and among all contributors in 2021 in Table 14-2, we noticed that, in most ecosystems the percentage among core contributors is higher than that among all contributors, with few exceptions such as Meteor, Pub, Cargo, and Hex, which have very small number of women contributors overall.

Table 14-2. *Women's participation by package managers (sorted by the number of projects)*

Ecosystem	Programming Language	# of Projects	% Women (2021)	Max % of Women	Window of the Max Pct	% Core Women (2021)	Max % Core Women
npm	JavaScript	568,116	5.36%	5.39%	Apr–Jun 2019	5.83%	5.83%
Packagist	PHP	250,687	3.23%	3.58%	Apr–Jun 2018	3.42%	3.89%
Go	Go	236,902	4.33%	4.59%	Oct–Dec 2019	4.49%	4.84%
PyPI	Python	116,819	5.33%	5.61%	Jan–Mar 2019	5.78%	6.03%
RubyGems	Ruby	94,561	5.7%	5.77%	Jul–Sep 2020	6.17%	6.24%
Bower	CSS	57,885	5.48%	5.48%	Jan–Mar 2021	5.76%	5.76%
CocoaPods	Objective-C	52,109	4.5%	4.85%	Oct–Dec 2018	4.66%	4.94%

(*continued*)

Table 14-2. (*continued*)

Ecosystem	Programming Language	# of Projects	% Women (2021)	Max % of Women	Window of the Max Pct	% Core Women (2021)	Max % Core Women
NuGet	C#	44,283	4.01%	4.01%	Jan–Mar 2021	4.63%	4.63%
Maven	Java	29,187	5.3%	5.36%	Apr–Jun 2019	5.84%	5.84%
Cargo	Rust	18,466	3.87%	4.52%	Apr–Jun 2014	3.65%	4.66%
Clojars	Clojure	12,551	4.79%	4.95%	Jul–Sep 2020	5.33%	5.33%
Atom	CSS	10,685	4.51%	5.82%	Jul–Sep 2019	5.75%	6.8%
CPAN	Perl	10,365	1.37%	6.15%	Jan–Mar 2008	2.04%	5.26%
Hex	Elixir	7,821	3.81%	3.81%	Jan–Mar 2021	3.64%	3.82%
Meteor	JavaScript	7,795	6.93%	6.93%	Jan–Mar 2021	6.25%	6.25%
Hackage	Haskell	7,570	3.4%	4.05%	Jan–Mar 2019	3.80%	4.09%
Pub	Dart	6,355	3.88%	6.25%	Oct–Dec 2012	3.74%	7.69%
CRAN	R	5,322	10.02%	10.02%	Jan–Mar 2021	10.51%	10.51%
Puppet	Puppet	3,943	1.49%	3.87%	Oct–Dec 2017	1.59%	4.15%
PlatformIO	C++	3,637	1.74%	4.55%	Apr–Jun 2012	2.63%	4.28%
Others	-	23,021					

Main Takeaways

The gender diversity is improving. We observed a slow but steadily increasing trend of women's participation in open source infrastructural projects. Our observation agrees with prior findings [41, 48]. The increasing trend is also observed in most of the ecosystems. While the reasons behind this change over time are beyond the scope of our study, we speculate that some of the past efforts to encourage and support marginalized groups in OSS have taken effect.

Gender distributions vary across ecosystems. Specifically, many ecosystems related to web development, especially front end, for example, Meteor and RubyGems, have higher women percentages. In comparison, several ecosystems related to system programming, for example, CPAN and PlatformIO, have lower gender diversity. Our finding agrees with Vasarhelyi et al.'s finding [56] that contributors in front-end programming languages are more likely to be women.

There are more core women contributors among big open source projects. When computing women's percentage among core contributors, we focused on only the biggest projects, whose commits are ranked top 10% in that ecosystem. We found that, among the biggest projects, whose commits are ranked the top 10% in that ecosystem, the percentage of women among core contributors is higher than that of among all contributors.

Open Research Questions

Reasons behind the increase: While our analysis and several recent studies [41, 48] reported a similar trend of increasing percentage of women among open source contributors, we do not yet understand how this has happened. Is it by chance or because some prior diversity efforts have been effective? Are hackathons [39], coding camps [43], or conferences effective in attracting and retaining women contributors? Future research can analyze the reasons behind the increased women's percentage and reflect on the outcome of prior efforts to improve diversity. Such studies can inform the design and deployment of future diversity and inclusion activities.

Ecosystem difference: Our study provides another piece of evidence that the differences in gender representation could be due to the functions of the programming languages. However, more in-depth and targeted studies are needed to test the speculation or provide a reasonable explanation. Is the disparity due to the nature of the programming languages or some community practices?

A fine-grained examination on women's representation across open source: Although our analysis found differences in gender representation across ecosystems and the level of contributions, there are more ways to slide the data and pinpoint the places with skewer gender distribution. For example, we examined the percentage of women core contributors among big projects and found that the percentage is higher than among all contributors. This is different from a prior result where the percentage of women among core contributors is much lower than that among all contributors [8]. Future studies can further investigate the relationship between gender distributions and project sizes. Our study also did not investigate the non-code contributions. Future researchers can consider adding contributors who only contributed to issue discussions. There are also non-code contributions that are not visible on social coding platforms. Quantifying the gender distribution among these hidden contributors is an open research question.

Bibliography

[1] Shaosong Ou and Alexander Hars. Working for free? Motivations for participating in open-source projects. *International Journal of Electronic Commerce*, 6(3):25–39, 2002.

[2] Luís A. Alexandre. Gender recognition: A multiscale decision fusion approach. *Pattern Recognition Letters*, 31(11):1422–1427, 2010.

[3] Ikram El Asri and Noureddine Kerzazi. Where are females in OSS projects? Socio technical interactions. In *Working Conference on Virtual Enterprises*, 308–319, Springer, 2019.

[4] Guilherme Avelino, Eleni Constantinou, Marco Tulio Valente, and Alexander Serebrenik. On the abandonment and survival of open source projects: An empirical investigation. In *International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–12, IEEE, 2019.

[5] Jelke Bethlehem. Selection bias in web surveys. *International Statistical Review*, 78(2):161–188, 2010.

[6] Christopher Bogart, Christian Kästner, James Herbsleb, and Ferdian Thung. How to break an API: cost negotiation and community values in three software ecosystems. In *Proceedings of the 2016 24th ACM*

SIGSOFT International Symposium on Foundations of Software Engineering, 109–120, 2016.

[7] Amiangshu Bosu and Kazi Zakia Sultana. Diversity and inclusion in open source software (OSS) projects: Where do we stand? In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–11, IEEE, 2019.

[8] Edna Dias Canedo, Rodrigo Bonifácio, Márcio Vinicius Okimoto, Alexander Serebrenik, Gustavo Pinto, and Eduardo Monteiro. Work practices and perceptions from women core developers in OSS communities. In *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–11, 2020.

[9] Elian Carsenat. Inferring gender from names in any region, language, or alphabet. *Unpublished*, 10, 2019.

[10] Hilary Carter and Jessica Groopman. The Linux Foundation report on diversity, equity, and inclusion in open source. https://www.linuxfoundation.org/tools/the-2021-linux-foundation-report-on-diversity-equity-and-inclusion-in-open-source/, 2021. Accessed on March 10, 2022.

[11] Gemma Catolino, Fabio Palomba, Damian A. Tamburri, Alexander Serebrenik, and Filomena Ferrucci. Gender diversity and women in software teams: How do they affect community smells? In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, 11–20, IEEE, 2019.

[12] Jailton Coelho and Marco Tulio Valente. Why modern open source projects fail. In *Proceedings of the Joint Meeting on Foundations of Software Engineering (ESEC/FSE)*, 186–196, ACM, 2017.

[13] Malcolm Corney, Olivier De Vel, Alison Anderson, and George Mohay. Gender-preferential text mining of e-mail discourse. In *18th Annual Computer Security Applications Conference, 2002, Proceedings.*, 282–289, IEEE, 2002.

[14] Daniel Izquierdo Cortázar. Gender-diversity analysis of the Linux kernel technical contributions. https://speakerdeck.com/bitergia/gender-diversity-analysis-of-the-linux-kernel-technical-contributions?slide=48, 2016. Accessed on January 20, 2022.

[15] Paul A. David, Andrew Waterman, and Seema Arora. Floss-us the free/libre/open source software survey for 2003. *Stanford Institute for Economic Policy Research, Stanford University, Stanford, CA (www.stanford.edu/group/floss-us/report/FLOSS-US-Report.pdf)*, 2003.

[16] Tapajit Dey, Sara Mousavi, Eduardo Ponce, Tanner Fry, Bogdan Vasilescu, Anna Filippova, and Audris Mockus. *Detecting and Characterizing Bots That Commit Code*, 209–219. ACM, New York, NY, USA, 2020.

[17] Edna Dias Canedo, Heloise Acco Tives, Madianita Bogo Marioti, Fabiano Fagundes, and José Antonio Siqueira de Cerqueira. Barriers faced by women in software development projects. *Information*, 10(10):309, 2019.

[18] Nadia Eghbal. *Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure*. Ford Foundation, 2016.

[19] Hongbo Fang, Daniel Klug, Hemank Lamba, James Herbsleb, and Bogdan Vasilescu. Need for tweet: How open source developers talk about their GitHub work on Twitter. In *Proceedings of the 17th International Conference on Mining Software Repositories*, 322–326, 2020.

[20] Sharan Foga. ASF committer diversity survey. https://cwiki.apache.org/confluence/display/COMDEV/ASF+Committer+Diversity+Survey+-+2016, 2016. Accessed on January 20, 2022.

[21] Denae Ford, Alisse Harkins, and Chris Parnin. Someone like me: How does peer parity influence participation of women on stack overflow? In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 239–243, IEEE, 2017.

[22] Marco Gerosa, Igor Wiese, Bianca Trinkenreich, Georg Link, Gregorio Robles, Christoph Treude, Igor Steinmacher, and Anita Sarma. The shifting sands of motivation: Revisiting what drives contributors in open source. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, 1046–1058, IEEE, 2021.

[23] GitHub. Open source survey. https://opensourcesurvey.org/2017/, 2017. Accessed on March 10, 2022.

[24] Rishab A. Ghosh, Ruediger Glott, Bernhard Krieger, and Gregorio Robles. Free/libre and open source software: Survey and study, 2002.

[25] Georgios Gousios and Diomidis Spinellis. GHTorrent: GitHub's data from a firehose. In *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*, 12–21, IEEE, 2012.

[26] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13, 2018.

[27] Daniel Izquierdo, Nicole Huesman, Alexander Serebrenik, and Gregorio Robles. OpenStack gender diversity report. *IEEE Software*, 36(1):28–33, 2018.

[28] Mitchell Joblin, Sven Apel, Claus Hunsen, and Wolfgang Mauerer. Classifying developers into core and peripheral: An empirical study on count and network metrics. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 164–174, IEEE, 2017.

[29] Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(*CSCW*):1–22, 2018.

[30] Andrew Kofink. Contributions of the under-appreciated: Gender bias in an open-source ecology. In *Companion Proceedings of the 2015 ACM SIGPLAN International Conference on Systems, Programming, Languages and Applications: Software for Humanity*, 83–84, 2015.

[31] Victor Kuechler, Claire Gilbertson, and Carlos Jensen. Gender differences in early free and open source software joining process. In *IFIP International Conference on Open Source Systems*, 78–93, Springer, 2012.

[32] Karim R. Lakhani and Robert G. Wolf. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Open Source Software Projects (September 2003)*, 2003.

[33] Amanda Lee and Jeffrey C Carver. Floss participants' perceptions about gender and inclusiveness: a survey. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, 677–687, IEEE, 2019.

[34] Feng Lin, Yingxiao Wu, Yan Zhuang, Xi Long, and Wenyao Xu. Human gender classification: a review. *Int. J. Biom.*, 8(3/4):275–300, 2016.

[35] Jeffrey W. Lockhart, Molly M King, and Christin Munsch. What' s in a name? Name-based demographic inference and the unequal distribution of misrecognition. 2022.

[36] Mike Medeiros, Benjamin Forest, and Patrik Öhberg. The case for non-binary gender questions in surveys. *PS: Political Science & Politics*, 53(1):128–135, 2020.

[37] Courtney Miller, David Widder, Christian Kästner, and Bogdan Vasilescu. Why do people give up FLOSSing? A study of contributor disengagement in open source. In *International Conference on Open Source Systems, OSS*, 116–129, Springer, 2019.

[38] Dawn Nafus. "Patches don't have gender": What is not open in open source software. *New Media & Society*, 14(4):669–683, 2012.

[39] Lavinia Paganini and Kiev Gama. Engaging women's participation in hackathons: A qualitative study with participants of a female-focused hackathon. In *International Conference on Game Jams, Hackathons and Game Creation Events 2020*, 8–15, 2020.

[40] Antoine Pietri, Diomidis Spinellis, and Stefano Zacchiroli. The software heritage graph dataset: public software development under one roof. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 138–142, IEEE, 2019.

[41] Gede Artha Azriadi Prana, Denae Ford, Ayushi Rastogi, David Lo, Rahul Purandare, and Nachiappan Nagappan. Including everyone, everywhere: Understanding opportunities and challenges of geographic gender-inclusion in OSS. *IEEE Transactions on Software Engineering*, 2021.

[42] Huilian Sophie Qiu, Alexander Nolte, Anita Brown, Alexander Serebrenik, and Bogdan Vasilescu. Going farther together: The impact of social capital on sustained participation in open source. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, 688–699, IEEE, 2019.

[43] Huilian Sophie Qiu, Yang Wen, and Alexander Nolte. Approaches to diversifying the programmer community – the case of the girls coding day. In *2021 IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, 91–100, IEEE, 2021.

[44] Huilian Sophie Qiu, Zihe H. (co-first author) Zhao, Tielin Katy Yu, Justin Wang, Alexander Ma, Hongbo Fang, Laura Dabbish, and Bogdan Vasilescu. Gender representation among contributors to open-source infrastructure – an analysis of 20 package manager ecosystems. In *International Conference on Software Engineering – Software Engineering in Society*, *ICSE SEIS*, IEEE, 2023.

[45] 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):1–25, 2017.

[46] 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 *IFIP International Conference on Open Source Systems*, 163–173, Springer, 2016.

[47] Gregorio Robles, Hendrik Scheider, Ingo Tretkowski, and Niels Weber. Who is doing it. *A Research on Libre Software Developers*, 2001.

[48] Davide Rossi and Stefano Zacchiroli. Worldwide gender differences in public code contributions: and how they have been affected by the COVID-19 pandemic. *Proceedings of the 44th International Conference on Software Engineering (ICSE 2022) – Software Engineering in Society (SEIS) Track*, 2022.

[49] Daniel Russo and Klaas-Jan Stol. Gender differences in personality traits of software engineers. *IEEE Transactions on Software Engineering*, 2020.

[50] Lucía Santamaría and Helena Mihaljević. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156, 2018.

[51] Paul Sebo. Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association: JMLA*, 109(3):414, 2021.

[52] Stack Overflow. Developer survey results. https://insights.stackoverflow.com/survey/ 2017, 2017. Accessed on May 1, 2022.

[53] 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 Comp Sci*, 3:e111, 2017.

[54] Bianca Trinkenreich, Mariam Guizani, Igor Wiese, Anita Sarma, and Igor Steinmacher. Hidden figures: Roles and pathways of successful OSS contributors. *Proceedings of the ACM on Human-Computer Interaction*, 4(*CSCW2*):1–22, 2020.

[55] Bianca Trinkenreich, Igor Wiese, Anita Sarma, Marco Gerosa, and Igor Steinmacher. Women's participation in open source software: A survey of the literature. Preprint at *arXiv:2105.08777*, 2021.

[56] Orsolya Vasarhelyi and Balazs Vedres. Gender typicality of behavior predicts success on creative platforms. Preprint at *arXiv:2103.01093*, 2021.

[57] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study of Stack Overflow. In *2012 International Conference on Social Informatics*, 332–338, IEEE, 2012.

[58] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study. *Interacting with Computers*, 26(5):488–511, 2014.

[59] 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 *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3789–3798, 2015.

[60] Bogdan Vasilescu, Alexander Serebrenik, and Vladimir Filkov. A data set for social diversity studies of GitHub teams. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, 514–517, IEEE, 2015.

[61] Bogdan Vasilescu, Alexander Serebrenik, Mathieu Goeminne, and Tom Mens. On the variation and specialisation of workload – a case study of the gnome ecosystem community. *Empirical Software Engineering*, 19(4):955–1008, 2014.

[62] Ji Zheng and Bao-Liang Lu. A support vector machine classifier with automatic confidence and its application to gender classification. *Neurocomputing*, 74(11):1926–1935, 2011.