Gender Representation Among Contributors to Open-Source Infrastructure

An Analysis of 20 Package Manager Ecosystems

Huilian Sophie Qiu*‡§, Zihe H Zhao†§, Tielin Katy Yu*, Justin Wang*,
Alexander Ma*, Hongbo Fang*, Laura Dabbish*, and Bogdan Vasilescu*

*Carnegie Mellon University, †Rice University, ‡Northwestern University
sophie.qiu@kellogg.northwestern.edu, zz83@rice.edu, {tieliny, justinw1, ajm3, hongbofa, dabbish, bogdanv}@andrew.cmu.edu

***Abstract*—While the severe underrepresentation of women and non-binary people in open source is widely recognized, there is little empirical data on how the situation has changed over time and which subcommunities have been more effectively reducing the gender imbalance. To obtain a clearer image of gender representation in open source, we compiled and synthesized existing empirical data from the literature, and computed historical trends in the representation of women across 20 open source ecosystems. While inherently limited by the ability of automatic name-based gender inference to capture true gender identities at an individual level, our census still provides valuable population-level insights. Across all and in most ecosystems, we observed a promising upward trend in the percentage of women among both core and general code contributors over time, but also high variation in the percentage of women contributors across ecosystems. We also found that, in most ecosystems, women withdraw earlier from open-source participation than men.**

***General Abstract*—The representation of women and non-binary people has been extremely low in the open-source software community. Most of the statistics reported by prior studies are below 10%. However, the majority of the prior works were based on subsamples instead of the entire population. Our work started with a review of the gender distributions reported in the literature. Then we provided an overview of the gender distribution in 20 of the largest open-source ecosystem, *i.e.*, grouped by package managers such as npm and PyPI, and investigated its change over time. Moreover, we analyzed the turnover rate between men and women contributors. Across all and in most ecosystems, we observed a promising upward trend in the percentage of women among both core and general code contributors over time, but also high variation in the percentage of women contributors across ecosystems. We also found that, in most ecosystems, women withdraw earlier from open-source participation than men.**

***Index Terms*—open-source software, gender diversity**

I. Introduction

It is a well-known phenomenon that the percentage of women contributors is low in the open-source software (OSS) community: various studies report this number being smaller than 10%. Low gender diversity is particularly problematic, as it hinders benefits a team could have possessed otherwise [1]–[3]. In attempting to address the low gender diversity problem, we first need a comprehensive understanding of the status quo of gender distribution. However, with few exceptions [4], [5], we lack a global overview of the gender distribution in OSS. Moreover, because most of the prior studies were based on different periods of time using different methods and sub-populations, the reported data vary significantly from one to the other. Our first contribution is a literature review of these prior works to display the wide variety of reported gender distributions along with their methods, periods, and samples. This survey provides a sense of disagreement among the studies and motivates the need to analyze at the community level.

Our second contribution is to provide an overview of the diversity status quo. We provide another data point with a further investigation of the gender distribution in OSS at the scale of the entire population. More importantly, our data focus on digital infrastructure projects with explicit open-source licenses instead of all public code contributions. These OSS projects are built to be reused, many of which are frequently utilized by open-source and commercial software developers to construct their packages and apps, constituting the foundation for much of our digital economy [6].

Moreover, although some studies looked at a sample of open-source projects, there lacks a thorough breakdown of different ecosystems, *i.e.*, sub-communities formed by library dependencies such as PyPI and npm. Because each ecosystem has its own practices in managing projects and contributions [7]–[9], it is reasonable to hypothesize that the gender distribution varies. Thus, our analysis further breaks down the data into ecosystems and examines the gender composition and disengagement of women contributors.

When investigating gender distribution, we followed many previous studies [4], [10] 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 [11]. We are aware that the use of the inference on individuals can be harmful [12], [13]. *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* [14].

In summary, our study starts with a literature review on gender distribution in open-source reported by prior studies. Then we provide another data point that enriches prior studies by answering the following research questions:

RQ$_1$. What is the gender distribution in OSS libraries overall? How does it change over time?

RQ$_2$. How does the gender representation in contributions vary across OSS ecosystems? How do they change over time?

RQ$_3$. How does the turnover rate of women contributors vary across ecosystems, compared with men contributors?

§Qiu and Zhao are co-first authors. Qiu performed the work while a Ph.D. Candidate at Carnegie Mellon University.

TABLE I: Women ratios reported from survey data.

Year	Source	Sample size	Ratio	Citation
2001	Online survey	5,478	0%	Robles *et al.* [15]
2002	Online survey	2,784	1.1%	Ghosh [16]
2001-2002	Email	684	2.5%	Lakhani *et al.* [17]
2002	Email	79	5%	Hars and Ou [18]
2003	Online Survey	1,588	1.6%	David *et al.* [19]
2013	Online survey	2,183	10.35%	Robles *et al.* [2]
2015	Online survey	816	24%	Vasilescu *et al.* [1]
2017	Online survey	6,000	5%	GITHUB [20]
2017	Online survey	64,000	7.6%	StackOverflow [21]
2019	Online survey	119	10.9%	Lee *et al.* [22]
2021	Online survey	242	7.6%	Gerosa *et al.* [23]

II. RELATED WORK

The main concern of this short paper is the data on open-source community's gender distribution. Therefore, this section focuses only on the data reported by prior studies without discussing the studies' details or findings regarding diversity problems or practices.

A. Prior reports on women percentage

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%. Here we provide an overview of the results reported by prior studies as a reference. The search for prior studies takes on a snowball sampling strategy: We started with the most recent works that reported gender distributions. Then we went through their references to find other studies that have reported gender distributions until we could no longer find papers that we did not cover.

We group prior works by their data collection methods. Note that since almost all the prior works relied on samples, they do not reflect longitudinal changes.

Surveys. Table I 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 [24]. However, survey data, albeit more reliable and accurate, are prone to selection bias [25]. Moreover, survey samples are usually small, making it hard to generalize.

Mining data. Table II 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, large-scale sample than the survey approach. It also eliminates the burden on the survey respondents and the efforts taken to collect survey results.

Ecosystems. Table III lists studies that report gender ratios in specific software ecosystems. The percentages of women range from 0% (Whamcloud) to 10% (OmapZoom) [32]. However, to the best of our knowledge, there is not a study that covers all major ecosystems, and many of these previous studies focus on specific projects rather than the entire ecosystem.

Geolocation. Recent studies began to look at the geolocation diversity among open-source contributors. Two studies [4], [5] found that, globally, gender diversity is low but has been increasing. The percentages of women vary on different continents.

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

[2]http://www.genderize.io

TABLE II: Women ratios reported from mining data.

Year	Source	Sample size	Ratio	Citation
2012	Email subs+US Census	1,931	8.27%	Kuechler *et al.* [26]
2012	SO	2,588	11.24%	Vasilescu *et al.* [10]
2015	GH+gC	1,049,345	8.71%	Kofink [27]
2015	GH+gC	873,392	9%	Vasilescu *et al.* [28]
2017	GH+social media	328,988	6.36%	Terrell *et al.* [29]
2017	OpenStack+genderize	-	10.4%	Izquierdo *et al.* [30]
2019	GH+Ns	300,000	9.7%	Qiu *et al.* [31]
2019	Gerrit+gC+social media	4,543	8.8%	Bosu and Sultana [32]
2020	GH+gC+Ns	1,954 core	5.35%	Canedo [33]
2021	GH+gC+SG	1,634,373	5.49%	Vasarhelyi *et al.* [34]
2021	GH+genderize	65,132	10%	Prana *et al.* [5]
2022	SH+GG	21.4M	10%	Rossi *et al.* [4]

Email subs: Email subscribers;
GH: GITHUB; SO: StackOverflow; SH: Software Heritage [35];
gC: genderComputer [10]; Ns:Namsor [36];
GG:GENDER GUESSER;[1] genderize: genderize.io;[2]
SG: SIMPLE GENDER [37]

TABLE III: Women ratios in different ecosystems.

Year	Source	Ecosystem	Sample size	Ratio	Citation
2014	Mailing list	Drupal	3,342	9.81%	Vasilescu *et al.* [38]
2014	Mailing list	Wordpress	3,611	7.81%	Vasilescu *et al.* [38]
2016	Online survey	Apache	765	5.2%	Sharan [39]
2005-16	GH	Linux	14,905	8%	Cortázar [40]
2016	Online survey	Debian	1,479	2%	Raissi *et al.* [41]
2019	GH+Ns	Angular.js	1,601	3.4%	Asri and Kerzazi [42]
2019	GH+Ns	Moby	1,824	3.5%	Asri and Kerzazi [42]
2019	GH+Ns	Rails	3,723	4.2%	Asri and Kerzazi [42]
2019	GH+Ns	Django	1,672	5.3%	Asri and Kerzazi [42]
2019	GH+Ns	Elasticsearch	1,127	4.2%	Asri and Kerzazi [42]
2019	GH+Ns	TensorFlow	1,735	5.8%	Asri and Kerzazi [42]
2019	Gerrit+gC	Android	258 core	3.87%	Bosu and Sultana [32]
2019	Gerrit+gC	Chromium OS	151 core	3.97%	Bosu and Sultana [32]
2019	Gerrit+gC	Couchbase	24 core	4.17%	Bosu and Sultana [32]
2019	Gerrit+gC	Go	90 core	7.77%	Bosu and Sultana [32]
2019	Gerrit+gC	LibreOffice	68 core	1.47%	Bosu and Sultana [32]
2019	Gerrit+gC	OmapZoom	60 core	10%	Bosu and Sultana [32]
2019	Gerrit+gC	oVirt	34 core	2.94%	Bosu and Sultana [32]
2019	Gerrit+gC	Qt	159 core	3.12%	Bosu and Sultana [32]
2019	Gerrit+gC	Typo3	73 core	4.1%	Bosu and Sultana [32]
2019	Gerrit+gC	Whamcloud	19 core	0%	Bosu and Sultana [32]
2021	Online survey	Linux	2,350	14%	Carter *et al.* [43]

GH: GITHUB; Ns: Namsor [36]; gC: genderComputer [10]

B. Contributors' turnover rate and gender

Turnover refers to team members disengaging from a project [44]. Given a constant need for efforts to develop and maintain open-source software, a high turnover rate harms projects' sustainability. The turnover rate, or survival rate, in open-source is high. Sharma *et al.* [45] found that, for a typical developer in a typical OSS project, the chance of turnover is 38% in 2 months. Turnover can be induced by various factors, such as personal expectation [46], project dissatisfaction [46], and low organizational commitment [47]. Qiu *et al.* [31] discovered that, in general, women disengage from open-source development earlier than men.

Many prior works focused on the turnover of marginalized groups, such as women or newcomers [31], and studied what can help improve their retention [48]. However, there lacks a breakdown of the turnover rate in different systems, which can be useful information when identifying how different management and practices affect the disengagement rate. Our study uses survival analysis to report the turnover rate between genders and provide a breakdown that can guide future researchers for more focused studies.

III. 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,[3] 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 GHTorrent [49], 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.

A. Data processing pipeline

Extracting the list of OSS 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.

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 [50], [51]. 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 *e.g.*, emails and names [52], [53].[4] For example, if two accounts use the same email and similar names, *i.e.*, some or all parts are the same but in different orders, or the same name with similar emails, *i.e.*, 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 [54]. 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 that window as no activity, as the percentage of women might surge and become an outlier in the data.

B. 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 [36], one of the name-based gender inference tools with the highest accuracy [11], [55]. 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.

C. Survival analysis

We use survival analysis, a statistical modeling technique for time-to-event data [56], to study the turnover rate. In our case, the event is a user's last commit in a particular ecosystem. For each user, we mark the three-month window within which they made the last commit as the time of their disengagement. For contributors whose last commit is later than March 2020, which is one year before the end of our data, we consider them to still be active at the time we collected data.

For each contributor, we denote the number of three-month windows between their first commit and last commit as their survival time T. A Kaplan-Meier curve plots the survival function $S(t)$ against time t, *i.e.*, a three-month window, to visualize the percentage of male or female contributors left after being active on GitHub for that many three-window months.

We run the Cox proportional hazards model [57] to assess the effect size of the gender factor on a contributor's survival rate. The Cox model is a semi-parametric regression that can estimate the effect of independent variables `gender`'s hazard ratio on the outcome variable compared to the baseline `being a man`. A hazard ratio greater than 1 means that `being a woman` is associated with a higher hazard rate, *i.e.*, the length of survival decreases.

IV. Results

This section presents our census findings. We summarize the important statistics and findings in Table IV for a quick overview. To reduce the effect of *Unknown* gender on our result, we calculate women fraction by

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

TABLE IV: Quick facts from census results

The largest ecosystem by projects/contributors	npm/npm
Ecosystem with the highest women percentage	CRAN (10.02%)
Ecosystem with the lowest women percentage	C
Women's % among all contributors in 2008	2.25%
Women's % among all contributors in 2021	4.87%
Ecosystem with the highest hazard ratio	PyPI

A. Gender composition in OSS libraries

Figure 1 shows the overall gender distribution in OSS libraries and its evolution over time, answering **RQ**$_1$. Overall, the percentage of women has been constantly low – no higher than 5.0%.

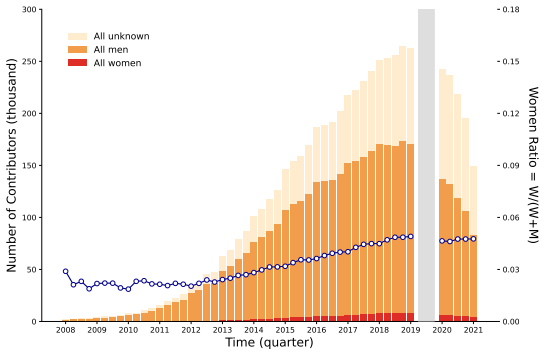
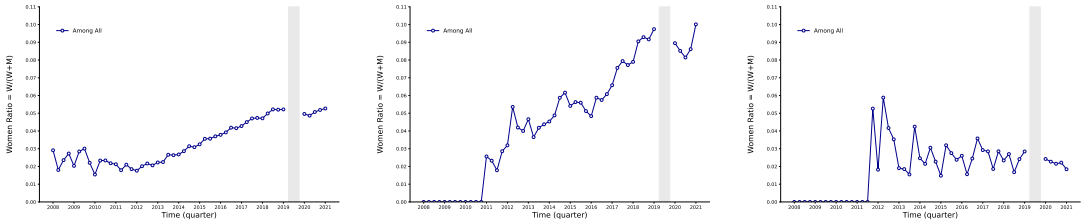


Fig. 1: Gender representation in OSS contribution overall. Grey bar covers the period where GHTorrent has missing data.

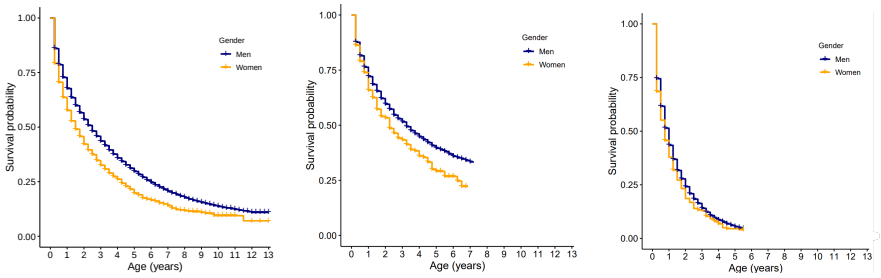
[3] https://libraries.io

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



(a) npm (JavaScript) (b) CRAN (R) (c) PlatformIO (C)

Fig. 2: Women distributions overall and in selected ecosystems. Grey bar covers the period with missing data on GHTorrent.



(a) PyPI (Python) (b) CRAN (R) (c) Meteor (JavaScript)

Fig. 3: Kaplan-Meier curves for selected ecosystems.

B. Gender distributions in different ecosystems

Regarding **RQ**$_2$, 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 three more representative ecosystems in Figure 2: npm, CRAN, and PlatformIO. For more figures, please visit our GitHub page.[5]

Figure 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 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 [7].

Moreover, as shown in Figure 2c, PlatformIO displays 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.

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

C. Turnover rate

Now we present the turnover rate of women contributors in different ecosystems, answering **RQ**$_3$. Figure 3 displays the three more representative survival plots, showing the Kaplan-Meier curves.

PyPI ecosystem (Figure 3a) has the highest hazard ratio of being a woman. The gap between the two curves keeps widening until year 6, and then it shrinks. This pattern suggests that, although women are leaving at a faster rate than men, their attrition rate becomes lower once they reach a certain length of tenure. Also, note that both curves have a non-zero survival rate towards the end of our dataset, meaning that PyPI is able to retain experienced contributors in both genders.

CRAN, having the highest percentage of women contributors, has a hazard ratio of 1.261, meaning that being a woman decreases the survival probability by 1.261 times. The two curves in CRAN (Figure 3b) keep widening, implying that women are leaving at a faster rate as they become more experienced in the field.

Figure 3c shows the two Kaplan-Meier curves of contributors to the Meteor ecosystem, the smaller package manager of the JavaScript libraries. Meteor's hazard ratio is much smaller than that of PyPI and closer to 1, which is reflected in the small gap between the two curves. Similar to PyPI, the two curves have a wider gap initially widens and later closes. Similar to CRAN, Meteor also lacks women contributors with long tenures. Table VI contains information for more ecosystems.

V. Discussions

This section mainly discusses speculations derived from our results and potential future research directions.

A. The gender diversity is improving

Observation: Answering **RQ**$_1$, we observed a slow but steadily increasing trend of women's participation in open-source infrastructural projects. Our observation agrees with prior findings [4], [5]. The increasing trend is also observed in most of the ecosystems (**RQ**$_2$).

Speculations and future directions: 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. Future research can analyze the specific reasons behind the increased women's percentage and reflect on the outcome of prior efforts to improve diversity.

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

TABLE V: Package managers and their corresponding main programming languages & women participation

Ecosystems	Language	# projects	% in 2021	Max %	Win of Max %
npm	JavaScript	568,116	5.36%	5.39%	Apr-Jun 2019
Packagist	PHP	250,687	3.23%	3.58%	Apr-Jun 2018
Go	Go	236,902	4.33%	4.59%	Oct-Dec 2019
PyPI	Python	116,819	5.33%	5.61%	Jan-Mar 2019
Rubygems	Ruby	94,561	5.7%	5.77%	Jul-Sep 2020
Bower	CSS	57,885	5.48%	5.48%	Jan-Mar 2021
CocoaPods	Objective-C	52,109	4.5%	4.85%	Oct-Dec 2018
NuGet	C#	44,283	4.01%	4.01%	Jan-Mar 2021
Maven	Java	29,187	5.3%	5.36%	Apr-Jun 2019
Cargo	Rust	18,466	3.87%	4.52%	Apr-Jun 2014
Clojars	Clojure	12,551	4.79%	4.95%	Jul-Sep 2020
Atom	CSS	10,685	4.51%	5.82%	Jul-Sep 2019
CPAN	Perl	10,365	1.37%	6.15%	Jan-Mar 2008
Hex	Elixir	7,821	3.81%	3.81%	Jan-Mar 2021
Meteor	JavaScript	7,795	6.93%	6.93%	Jan-Mar 2021
Hackage	Haskell	7,570	3.4%	4.05%	Jan-Mar 2019
Pub	Dart	6,355	3.88%	6.25%	Oct-Dec 2012
CRAN	R	5,322	10.02%	10.02%	Jan-Mar 2021
Puppet	Puppet	3,943	1.49%	3.87%	Oct-Dec 2017
PlatformIO	C++	3,637	1.74%	4.55%	Apr-Jun 2012
Others	-	23,021			

TABLE VI: Cox proportional hazards model results in 20 ecosystems, sorted by the coefficient of gender.

Ecosystem	# M	# W	Women Pct.	Hazard ratio	p-value
PyPI	94546	5115	0.051	1.355	0 ***
npm	221378	11082	0.048	1.341	0 ***
Packagist	71154	2417	0.033	1.313	0 ***
CocoaPods	29424	1289	0.042	1.302	0 ***
Bower	71834	2810	0.038	1.299	0 ***
Rubygems	58821	2778	0.045	1.289	0 ***
NuGet	41111	1399	0.033	1.275	0 ***
Maven	76573	3388	0.042	1.273	0 ***
CPAN	2040	78	0.037	1.261	0.069
CRAN	4015	408	0.092	1.261	0.001 **
Clojars	9260	328	0.034	1.258	0 ***
PlatformIO	2743	77	0.027	1.25	0.11
Go	53457	2299	0.041	1.249	0 ***
Puppet	4466	124	0.027	1.243	0.026 .
Cargo	9952	412	0.04	1.224	0.007 *
Meteor	8704	292	0.032	1.129	0.049 .
Hackage	3789	118	0.03	1.065	0.586
Pub	3960	147	0.036	1.048	0.705
Atom	4561	142	0.03	1.026	0.772
Hex	5145	158	0.03	0.981	0.849

Signif. Code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1" 1

B. Ecosystem difference

Observation: Answering $\mathbf{RQ}_2$, which investigates the gender distribution across different ecosystems, we found that gender distributions vary across ecosystems. Specifically, many ecosystems related to web development, especially front-end, *e.g.*, Meteor and Rubygems, have higher women percentages. In comparison, several ecosystems related to systems programming, *e.g.*, CPAN and Platform IO, have lower gender diversity.

Speculations and future directions: Our finding agrees with Vasarhelyi etal's finding [34] that contributors in front-end programming languages are more likely to be women. 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.

C. Higher turnover rate for women

Observation: In most ecosystems, women face a higher hazard ratio, meaning that being a woman is associated with a shorter survival length. Women's hazard ratios vary slightly across ecosystems.

Speculations and future directions: We do not yet have a clear speculation of what causes the differences in hazard ratio. It is possible that the differences are due to ecosystems' different natures, such as the functionality or dependency among projects [58]. Future work exploring this direction may require the application of social science theories, such as social capital [31], social network analysis, or organizational theories.

D. Presenting data

The goal of our study goes beyond computing gender distributions. We intend to publish the data to benefit a wider audience. We are building a website that can present the data we have collected in this paper, including the statistics from prior studies and results from our analysis. Moreover, we are compiling methods and tips from prior studies that can help open-source practitioners build a more diverse open-source community.

VI. Limitations

A. Gender inference method

The imperfect accuracy and binary gender assumption of our name-based gender inference method can lead to bias or intensify stereotype [59]. Moreover, while around 60% of the OSS contributors disclosed their names, we choose to retain gender labels that are classified with > 0.7 confidence for better accuracy, leaving us only 53.03% of the entire OSS population's gender identified. Future studies can work on a comprehensive census project that combines the usage of surveys and data mining, which can often provide more accurate gender data.

B. Maybe women hide their gender identity

Our inference is subject to a threat of systematic bias that women might be unwilling to disclose their gender to avoid discrimination. Although we do not have validated evidence for this speculation, we have heard about such practice from several women contributors.

C. Non-code contribution

We acknowledge that our methods neglected non-code contributions. Prior works discovered that many women contributors work on non-code contributions, such as community leader, coordinator, or administrator roles [50], and these community-centric roles are usually hidden behind GitHub's public traceable data [51]. Prana *et al.* [5] took an initial step in addressing this problem by collecting authors that did not make a commit but created or commented on issues. Future works can include activities from contributions such as issues or code review comments.

D. Missing data

Our plots show a slight drop in women's participation in 2020. Although this trend agrees with Rossi and Zacchiroli's study [4], the drop can also be caused by the fact that GHTorrent stopped collecting data after the first quarter of 2021.[6] Unfortunately, due to the large size of data in that time period, we were not able to retrieve all activities from GitHub directly.

VII. Conclusions

This paper presented a literature review on previously reported OSS gender distributions and a comprehensive analysis of women's participation across 20 OSS ecosystems. Overall, our results showed a slow yet steady increasing trend providing hope for future research and efforts to improve diversity in OSS. Based on our observations, we provided many speculations that could inspire further investigations.

[6]https://ghtorrent.org

References

[1] B. Vasilescu, D. Posnett, B. Ray, M. G. van den Brand, A. Serebrenik, P. Devanbu, and V. Filkov, "Gender and tenure diversity in github teams," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 3789–3798.

[2] G. Robles, L. A. Reina, J. M. González-Barahona, and S. D. Domínguez, "Women in free/libre/open source software: The situation in the 2010s," in *IFIP International Conference on Open Source Systems*. Springer, 2016, pp. 163–173.

[3] G. Catolino, F. Palomba, D. A. Tamburri, A. Serebrenik, and F. 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)*. IEEE, 2019, pp. 11–20.

[4] D. Rossi and S. 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.

[5] G. A. A. Prana, D. Ford, A. Rastogi, D. Lo, R. Purandare, and N. Nagappan, "Including everyone, everywhere: Understanding opportunities and challenges of geographic gender-inclusion in oss," *IEEE Transactions on Software Engineering*, 2021.

[6] N. Eghbal, *Roads and bridges: The unseen labor behind our digital infrastructure*. Ford Foundation, 2016.

[7] C. Bogart, C. Kästner, J. Herbsleb, and F. 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*, 2016, pp. 109–120.

[8] A. Mockus, R. T. Fielding, and J. D. Herbsleb, "Two case studies of open source software development: Apache and mozilla," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 11, no. 3, pp. 309–346, 2002.

[9] K. Nakakoji, Y. Yamamoto, Y. Nishinaka, K. Kishida, and Y. Ye, "Evolution patterns of open-source software systems and communities," in *Proceedings of the international workshop on Principles of software evolution*, 2002, pp. 76–85.

[10] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study of stackoverflow," in *2012 International Conference on Social Informatics*. IEEE, 2012, pp. 332–338.

[11] L. Santamaría and H. Mihaljević, "Comparison and benchmark of name-to-gender inference services," *PeerJ Computer Science*, vol. 4, p. e156, 2018.

[12] F. Hamidi, M. K. Scheuerman, and S. 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*, 2018, pp. 1–13.

[13] O. Keyes, "The misgendering machines: Trans/hci implications of automatic gender recognition," *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–22, 2018.

[14] J. W. Lockhart, M. M. King, and C. Munsch, "What's in a name? name-based demographic inference and the unequal distribution of misrecognition," 2022.

[15] G. Robles, H. Scheider, I. Tretkowski, and N. Weber, "Who is doing it," *A research on Libre Software developers*, 2001.

[16] R. A. Ghosh, R. Glott, B. Krieger, and G. Robles, "Free/libre and open source software: Survey and study," 2002.

[17] K. R. Lakhani and R. 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.

[18] S. O. Alexander Hars, "Working for free? motivations for participating in open-source projects," *International journal of electronic commerce*, vol. 6, no. 3, pp. 25–39, 2002.

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

[20] GitHub, "Open source survey," https://opensourcesurvey.org/2017/, 2017, accessed: 2022-03-10.

[21] StackOverflow, "Developer survey results," https://insights.stackoverflow.com/survey/2017, 2017, accessed: 2022-05-01.

[22] A. Lee and J. C. Carver, "Floss participants' perceptions about gender and inclusiveness: a survey," in *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 2019, pp. 677–687.

[23] M. Gerosa, I. Wiese, B. Trinkenreich, G. Link, G. Robles, C. Treude, I. Steinmacher, and A. Sarma, "The shifting sands of motivation: Revisiting what drives contributors in open source," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021, pp. 1046–1058.

[24] M. Medeiros, B. Forest, and P. Öhberg, "The case for non-binary gender questions in surveys," *PS: Political Science & Politics*, vol. 53, no. 1, pp. 128–135, 2020.

[25] J. Bethlehem, "Selection bias in web surveys," *International statistical review*, vol. 78, no. 2, pp. 161–188, 2010.

[26] V. Kuechler, C. Gilbertson, and C. Jensen, "Gender differences in early free and open source software joining process," in *IFIP International Conference on Open Source Systems*. Springer, 2012, pp. 78–93.

[27] A. 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*, 2015, pp. 83–84.

[28] B. Vasilescu, A. Serebrenik, and V. Filkov, "A data set for social diversity studies of github teams," in *2015 IEEE/ACM 12th working conference on mining software repositories*. IEEE, 2015, pp. 514–517.

[29] J. Terrell, A. Kofink, J. Middleton, C. Rainear, E. Murphy-Hill, C. Parnin, and J. Stallings, "Gender differences and bias in open source: Pull request acceptance of women versus men," *PeerJ Comp Sci*, vol. 3, p. e111, 2017.

[30] D. Izquierdo, N. Huesman, A. Serebrenik, and G. Robles, "Openstack gender diversity report," *IEEE Software*, vol. 36, no. 1, pp. 28–33, 2018.

[31] H. S. Qiu, A. Nolte, A. Brown, A. Serebrenik, and B. 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)*. IEEE, 2019, pp. 688–699.

[32] A. Bosu and K. Z. 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)*. IEEE, 2019, pp. 1–11.

[33] E. D. Canedo, R. Bonifácio, M. V. Okimoto, A. Serebrenik, G. Pinto, and E. 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)*, 2020, pp. 1–11.

[34] O. Vasarhelyi and B. Vedres, "Gender typicality of behavior predicts success on creative platforms," *arXiv preprint arXiv:2103.01093*, 2021.

[35] A. Pietri, D. Spinellis, and S. Zacchiroli, "The software heritage graph dataset: public software development under one roof," in *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, 2019, pp. 138–142.

[36] E. Carsenat, "Inferring gender from names in any region, language, or alphabet," *Unpublished*, vol. 10, 2019.

[37] D. Ford, A. Harkins, and C. 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)*. IEEE, 2017, pp. 239–243.

[38] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study," *Interacting with Computers*, vol. 26, no. 5, pp. 488–511, 2014.

[39] S. Foga, "Asf committer diversity survey. 2016," https://cwiki.apache.org/confluence/display/COMDEV/ASF+Committer+Diversity+Survey+-+2016, 2016, accessed: 2022-01-20.

[40] D. I. 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: 2022-01-20.

[41] M. Raissi, M. de Blanc, and S. Zacchiroli, "Preliminary report on the influence of capital in an ethical-modular project: Quantitative data from the 2016 debian survey," *Journal of Peer Production*, no. 10, pp. 1–25, 2017.

[42] I. E. Asri and N. Kerzazi, "Where are females in oss projects? socio technical interactions," in *Working Conference on Virtual Enterprises*. Springer, 2019, pp. 308–319.

[43] H. Carter and J. 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: 2022-03-10.

[44] M. Foucault, M. Palyart, X. Blanc, G. C. Murphy, and J.-R. Falleri, "Impact of developer turnover on quality in open-source software," in *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, 2015, pp. 829–841.

[45] P. N. Sharma, J. Hulland, and S. Daniel, "Examining turnover in open source software projects using logistic hierarchical linear modeling approach," in *IFIP International Conference on Open Source Systems*. Springer, 2012, pp. 331–337.

[46] Y. Yu, A. Benlian, and T. Hess, "An empirical study of volunteer members' perceived turnover in open source software projects," in *2012 45th Hawaii International Conference on System Sciences*. IEEE, 2012, pp. 3396–3405.

[47] P. Hynninen, A. Piri, and T. Niinimäki, "Off-site commitment and voluntary turnover in gsd projects," in *2010 5th IEEE International Conference on Global Software Engineering*. IEEE, 2010, pp. 145–154.

[48] B. Lin, G. Robles, and A. Serebrenik, "Developer turnover in global, industrial open source projects: Insights from applying survival analysis," in *2017 IEEE 12th International Conference on Global Software Engineering (ICGSE)*. IEEE, 2017, pp. 66–75.

[49] G. Gousios and D. Spinellis, "Ghtorrent: Github's data from a firehose," in *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*. IEEE, 2012, pp. 12–21.

[50] B. Trinkenreich, I. Wiese, A. Sarma, M. Gerosa, and I. Steinmacher, "Women's participation in open source software: A survey of the literature," *arXiv preprint arXiv:2105.08777*, 2021.

[51] B. Trinkenreich, M. Guizani, I. Wiese, A. Sarma, and I. Steinmacher, "Hidden figures: Roles and pathways of successful oss contributors," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–22, 2020.

[52] B. Vasilescu, A. Serebrenik, M. Goeminne, and T. Mens, "On the variation and specialisation of workload—a case study of the gnome ecosystem community," *Empirical Software Engineering*, vol. 19, no. 4, pp. 955–1008, 2014.

[53] H. Fang, D. Klug, H. Lamba, J. Herbsleb, and B. 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*, 2020, pp. 322–326.

[54] T. Dey, S. Mousavi, E. Ponce, T. Fry, B. Vasilescu, A. Filippova, and A. Mockus, *Detecting and Characterizing Bots That Commit Code*. New York, NY, USA: ACM, 2020, p. 209–219. [Online]. Available: https://doi.org/10.1145/3379597.3387478

[55] P. Sebo, "Performance of gender detection tools: a comparative study of name-to-gender inference services," *Journal of the Medical Library Association: JMLA*, vol. 109, no. 3, p. 414, 2021.

[56] R. G. Miller Jr, *Survival analysis*. Wiley, 2011, vol. 66.

[57] P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: a large sample study," *The annals of statistics*, pp. 1100–1120, 1982.

[58] M. Valiev, B. Vasilescu, and J. Herbsleb, "Ecosystem-level determinants of sustained activity in open-source projects: A case study of the pypi ecosystem," in *Proceedings of the Joint Meeting on Foundations of Software Engineering (ESEC/FSE)*. ACM, 2018, pp. 644–655.

[59] B. Custers, "Data dilemmas in the information society: Introduction and overview," in *Discrimination and privacy in the information society*. Springer, 2013, pp. 3–26.