Including Everyone, Everywhere: Understanding Opportunities and Challenges of Geographic Gender-Inclusion in OSS

Gede Artha Azriadi Prana, Denae Ford, Ayushi Rastogi, David Lo, Rahul Purandare, Nachiappan Nagappan

Abstract—The gender gap is a significant concern facing the software industry as the development becomes more geographically distributed. Widely shared reports indicate that gender differences may be specific to each country. However, how complete can these reports be with little to no research reflective of the Open Source Software (OSS) process and communities software is now commonly developed in? Our study presents a multi-region geographical analysis of gender inclusion on GitHub. This mixed-methods approach includes quantitatively investigating differences in gender inclusion in projects across geographic regions and investigate these trends over time using data from contributions to 21,456 project repositories. We also qualitatively understand the unique experiences of developers contributing to these projects through strategically targeted surveys. Our findings indicate that there are statistically significant differences in gender diversity between regions. Since 2014, there has been a small and statistically significant improvement of gender diversity among software project contributors in Northern America and South-Eastern Asia but negligible change elsewhere. We also find that most motivations and barriers to contributions (e.g. lack of resources to contribute and poor working environment) were shared across regions, however, some insightful differences, such as how to make projects more inclusive, did arise. From these findings, we derive and present implications for tools that can foster inclusion in open source software communities and empower contributions from everyone, everywhere.

Index Terms—inclusion, OSS, software engineering, empirical studies, GitHub, diversity, gender, geographic regions,

1 Introduction

Over the last few years there has been significant attention paid to the gender gap in the broader software industry. There are a number of efforts by organizations like Grace-Hopper[1], CRA-W[2], European Union's Women in Digital initiative[3], NCWIT[4], NASSCOM[5] and other to address this. There have also been efforts by the open source community to embrace gender diversity [1]. That said in its most recent report [2], NCWIT found out that more women are earning Associates, Bachelor's, Master's and Doctoral Computing Degrees, but across all these levels there is a persistent gender gap. Women degree earners are more racially diverse than men degree earners [2]. Further, Figure 1 below from the NCWIT report [2] shows the gender ratio as 1:3.03 in 2009 and 1:2.91 in 2017.

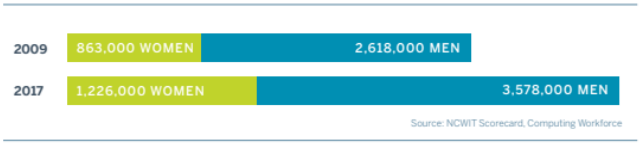

Fig. 1. The number of individuals in computing occupations (reproduced from [2]).

Correspondingly the report by the European Union [3] found out that only 24 out of every 1000 female tertiary graduates have a degree in an Information and communications technology (ICT) related subject—of which only six go on to work in the digital sector [3]. This was a decrease compared to 2011 [3]. The study also found that if more women were to enter the digital jobs market, it could create an annual EUR 16 billion GDP boost for the European economy [3]. Further, the study found that Finland, Sweden, Luxembourg, and Denmark have highest scores for participation of women in their digital economies compared to other states, whereas women are the least digital in Bulgaria, Romania, Greece, and Italy. Luxembourg, France, and Spain are more advanced in women in digital than their overall digital participation [3]. This brings up an interesting point that different countries have different circumstances for women in ICT.

Similarly, a reliable source for India is the NASSCOM report on the status of women in the IT industry in India [4]. This is a more detailed source which analyzes the distribution of women by career levels. The key findings in this report indicated women were concentrated at the lower career levels and there were fewer women in top position in IT companies in India. [4]. In addition, the report found that women also had a slower career progression compared to men and had higher retention issues [4].

The results across NCWIT, the EU and NASSCOM show significant differences according to the geography between the US, Europe and India.

This leads to our fundamental motivation to study if there are differences between gender diversity of software

1. https://ghc.anitab.org/
2. https://cra.org/cra-wp/
3. https://ec.europa.eu/digital-single-market/en/women-ict
4. https://www.ncwit.org/
5. https://www.nasscom.in/

project contributors in different parts of the world? Towards this end, the first and seminal piece of work in this area was done by Vasilescu et al. [5]. Using data from GitHub, authors studied how gender and tenure diversity relate to team productivity and turnover. They found that both gender and tenure diversity are positive and significant predictors of productivity, however they do not highlight distinctions across geographic regions nor the challenges these contributors face.

To better understand gender inclusion of open source software (OSS) across geographic regions, we conducted an empirical investigation of contributions on GitHub—a widely used social coding platform with over 50 million contributors [6]. In this study, we started with 125,485,095 GitHub projects and conducted a multi-stage filtering process based on project activity criteria to ensure we gathered a robust set of contributions styles. This filtering resulted in 21,456 projects being included in our analysis of project inclusion across gender and geographic region. To gather an account of interactions from developers of these projects, we strategically identified 1562 contributors across geographic regions and binary genders to distribute surveys to about their experiences. We found that although gender diversity has had a small statistical improvement over time, there is still room to create a more inclusive experience for OSS project contributors.

Our study of OSS projects builds on the work of Vasilescu et al. [5] by contributing: 1) a qualitative and quantitative perspective into gender inclusion of open source software projects across geographic regions, 2) an emerging taxonomy of motivations and challenges to contributing to gender inclusive OSS projects across geographic regions, 3) a publicly available dataset to encourage further investigations, and 4) empirically-informed design opportunities to make open source software projects more gender inclusive across geographic regions. In this paper, we investigate the following research questions:

RQ 1: Is there a difference between gender diversity of OSS software project contributors in different parts of the world? Gender has been studied independently [5] and so has geography [7], but not together. We ask this question to better understand the combination of these two factors as they are distinct constructs and have been noted to have different approaches to supporting interactions as reported in different technical career reports published in countries around the world (e.g., NCWIT, NASSCOM and EU). There has been no work to the best of our knowledge that addresses this research question.

RQ 2: Are there different rates of increase or decrease of gender diversity in OSS in different parts of the world? Understanding the trends of OSS communities of contributors over time can help us understand what the future of software development may look like in the future and indifferent parts of the world. This research question helps us define a baseline for others conducting analysis and research in OSS communities to be aware of how these trends may be different across different countries.

RQ 3: What are the factors motivating and inhibiting participation in OSS in different regions? How do they differ across regions? Previous work conducted to understand inclusive communities on GitHub have primarily focused on understanding the challenges of contributors [8]. Inevitably, most miss an opportunity to understand the perspectives of the marginalized in these communities. Better comprehension of the experiences of the marginalized, especially developers who are women and those who are from the 'Global South' [9], can allow us to build better interventions that foster inclusive developer experiences around the world.

2 Background

Success of open source software projects is attributed to its developers. This inspired a series of studies exploring reasons for open source engagement. These studies include motivations for developer participation [10], barriers to participation [11], and how developers contribute to open source [12]. These studies help understand and optimize the opportunities to retain community participation. It also prepares projects to avoid or mitigate situations that causes contributors to leave projects.

This paper is inspired by and extends works on barriers to participation in open source software projects along the lines of diversity in terms of gender and region of contributors in software projects. Next, we present some important research studies that have shaped the field.

Barriers to participation. Studies have shown that there are barriers that can prevent developers from participating to open source. These barriers can relate to social factors [11] and tools/technology [13], as well as the community it affects (e.g. newcomers [11], and generally under-represented communities [14]). These studies not only help in identifying the source of problem but also propose solutions to mitigate it [15].

Diversity in open source software projects. Diversity in open source software projects has gained widespread attention in recent years. Starting from the awareness of diversity and particularly the demographic attributes of developers [16], today improving diversity is seen as a goal for fairness [17] as well as improved productivity [5]. Many studies relating to gender diversity and the lack thereof followed discussing its relevance [17] and the impediments to improve gender diversity [18]. Along the similar lines, diversity in terms of nationality or region took off [19].

All these studies identify challenges and needs of under-represented communities. Taking research on the subject a step further, in this work we study gender diversity in different regions and how factors relating to gender and region can potentially explain why developers join open source software projects, select a project, continue participation. Such factors can potentially also explain barriers and reasons to leave a project.

3 Methodology

3.1 Overview of GitHub Dataset

To answer our research questions, we used the GHTorrent [20] dataset to identify GitHub users contributing to open source projects. We chose to use GHTorrent data as it has been widely used in software engineering research,

including in works related to diversity (e.g. [5], [17], [21]). Using the latest GHTorrent database dump (1 June 2019), we begin by filtering for repositories that are active, are not toy repositories, and involve collaboration between different developers. We use the following repository criteria:

- The repository has existed for at least 180 days (measured using difference of *updated_at* and *created_at* columns in the GHTorrent data).
- The repository has at least one commit from the beginning of 2018 or later.
- The repository has at least 10 commits from 4 or more distinct commit authors, none of which are marked 'fake' or 'deleted' in the GHTorrent user information table.
- The repository is not a fork.

We subsequently attempt to resolve the location and gender of the commit authors. As GHTorrent data does not include personal information, we collect additional information through the GitHub API prior to location and gender resolution. For location, resolution is based on value of *country_code* field of the commit author's user information, if available. If the field is empty, location resolution is attempted using the following:

- *location* field. For example, if the commit author specifies "Seattle" as his/her location, the country assigned will be USA. If he or she specifies "Tokyo", the country assigned will be "Japan".
- Latitude and longitude (*lat* and *long* fields in GHTorrent data, respectively).
- *company* field. For example, "Argonne National Lab" or "Puget Sound Regional Council" are considered as evidence that the commit author is based in the USA. "German National Library" is considered as evidence that the author is based in Germany. Where possible, we attempt to resolve an organization's location using its website and LinkedIn page. In case of multinational organizations, the author's location is considered unresolved unless more specific information such as branch name is provided. For example, "RedHat" will be considered as unresolved location, whereas "RedHat UK" will be considered as evidence that the location is the UK.
- *email* field. For example, if the author's email address uses an Australian government domain, the country assigned will be Australia.

To facilitate analysis at regional level, we also assign three levels of region information to each commit author based on the taxonomy of regions specified by United Nations Statistics Division[6]. For example, if the commit author's resolved location is Kenya, the assigned region information will be "Africa" (region level 1), "Sub-saharan Africa" (region level 2), and "Eastern Africa" (region level 3).

For the commit authors' genders, resolution is attempted by identifying first name portion of the commit author's name followed by resolution of gender using *genderize.io*[7], which is reported to have high accuracy [22] and has been used for similar purposes in other domains[8]. For this part, titles (e.g. "Dr.") are ignored, and if the commit author does not use Latin alphabet to specify their name, the name is first converted to Latin alphabet. As an additional measure to evaluate *genderize.io*'s accuracy, one of the authors randomly selected five sample repositories associated with different regions (two from Americas and one each from Asia, Europe, and Oceania). Each sample is then assigned to each of the remaining authors who subsequently attempt manual gender resolution using public information sources (the contributor's GitHub page, LinkedIn page, Twitter profile, etc.). The result is subsequently compared to gender prediction result from *genderize.io*. We find that the manual analysis results match genderize.io's results 89.5% of the time.

Following this, we apply further filtering for repositories for which both gender and location can be resolved for at least 75% of the commit authors. Considering that not all repositories on GitHub are software project repositories [23], we also exclude repositories for which GitHub detects no primary language. In all, after the entire process, 21,456 repositories are shortlisted, with the breakdown of filtering result at various stages shown in Table 1. We also extract all commit authors associated with the shortlisted repositories. Tables 2 and 3 show the statistics of the dataset.

TABLE 1
Result of project repository filtering steps.

Filtering step	Count
Initial number of repositories	125,485,095
Repositories with commits newer than January 1, 2018	31,947,039
Repositories that have existed for at least 180 days and are not marked as "deleted	4,393,507
Repositories with at least 10 commits, and are not a fork	2,129,448
Repositories remaining with no commit authors marked "fake" or "deleted"	97,989
Repositories with 75% commit authors having resolvable gender and location	21,456

TABLE 2
Statistics of shortlisted repositories.

	Min	Max	Mean	Median
No. of Contributors	4	109	6.16	5
No. of Commits	22	301692	363.27	170
Creation year	2008	2018	2014.63	2015

3.2 Protocol

To answer RQ1, we first associate a repository to a location based on the most common identified location of the commit authors. For example, if five commit authors contribute to a repository, and their locations are {"Germany", "USA", "USA", "USA", "Italy"}, then the repository will be associated with USA. We also apply the same processing at different region levels. Using the previous example, since the commit authors' level 1 regions are {"Europe", "Americas", "Americas", "Americas", "Europe"}, the repository

6. https://unstats.un.org/unsd/methodology/m49/
7. http://www.genderize.io
8. https://genderize.io/use-cases

TABLE 3
Commit author region and gender in shortlisted repositories.

Region Level 1	Region Level 2	Count	Percentage		
			Man	Woman	Unknown
Africa	Northern Africa	91	91.21	5.49	3.33
Africa	Sub-Saharan Africa	273	92.67	3.66	3.66
Americas	Latin America and the Caribbean	2547	93.29	4.75	1.96
Americas	Northern America	24055	90.27	7.47	2.25
Americas	Others	5	80.00	0.00	20.00
Asia	Central Asia	34	88.24	2.94	8.82
Asia	Eastern Asia	2585	80.46	10.10	9.44
Asia	South-eastern Asia	686	87.90	6.85	5.25
Asia	Southern Asia	1463	91.46	5.47	3.08
Asia	Western Asia	529	93.19	3.40	3.40
Europe	Eastern Europe	3858	94.35	2.90	2.75
Europe	Northern Europe	7541	92.71	5.38	1.91
Europe	Southern Europe	2314	94.77	3.11	2.12
Europe	Western Europe	10637	92.94	3.88	3.18
Oceania	Australia and New Zealand	1870	92.62	5.13	2.25
Oceania	Melanesia	5	80.00	0.00	20.00
Oceania	Polynesia	5	100.00	0.00	0.00
Unknown	Unknown	12123	61.96	6.22	31.82

will be associated with "Americas". We note that unlike the case with commit authors, there is possibility that in some cases, a repository may be associated with different parts of the world at different level. For example, a repository with commit author country set of {"Japan", "Singapore", "India", "Thailand", "USA", "USA", "USA"} will be associated with "USA" at country level (since "USA" is the country with the largest number of contributors), but "Asia" at region level 1 (since most contributors are from Asia). In view of this, during our analyses of the repositories, we consider each associated region level independently instead of as a hierarchy. We also focus our analyses at higher levels of regional grouping. We considered the alternative of associating repository to the location of the top contributor, however, this approach also has a potential issue if the contribution of the top commit author is outweighed by collective contribution from a different region. For example, if the top individual contributor is from the Americas and contributes 10% of the commits, but project members from Europe collectively contributes 50% and the remaining 40% is contributed by members in Asia, then the project will be associated with Americas, although most of the contribution comes from members in other parts of the world. Given our research objectives, we believe associating each project with the most common location of its contributors will be more appropriate than associating it with the location of its top contributor.

Afterwards, we measure geographic and gender diversity of the repositories using the Blau diversity index [24] which has also been used in several works in software engineering domain [5], [25], [26]. During calculation of indices for both type of diversity, we disregard unknown values. For example, if a repository is associated with five commit authors, and four of them are identified as men while one is unknown, the gender diversity index will be 0. Similarly, if a repository's set of commit authors comprise two men, two women, and one person with unidentified gender, the gender diversity index will be 0.5. We use the same approach for geographic diversity, i.e. if there are five commit contributors with the identified region level 1 of {"Americas","Americas","Asia","Asia","Unknown"}, the diversity index for region level 1 will be 0.5. To check for normality of the distribution of diversity indices, we apply D'Agostino's K^2 test [27] considering the large size of sample set. We subsequently apply Kruskal-Wallis H test [28] to detect statistically significant difference of diversity between regions, followed by Mann-Whitney U test [29] between pairs of regions, with Bonferroni correction [30] to reduce error due to multiple comparisons. After this test, we apply Cliff's Delta test [31] on pairs of regions with adjusted $p \leq 0.05$ to measure effect size[9]. Following [31], we define $|\ \delta\ |$ of less than 0.147 as negligible effect size, between 0.147 to 0.33 as small, between 0.33 and 0.474 as medium, and above 0.474 as large. We also computed Spearman's rho [32] to investigate correlation between gender diversity and geographic diversity at different levels of regional grouping. We use *SciPy* [33] implementation of these statistical tests.

To answer RQ2, we used the commit data from our GHTorrent dataset to reconstruct repositories' set of commit authors at an earlier point in time. We subsequently perform the same Blau diversity index analyses used to answer RQ1 to examine changes in diversity of repositories in the regions of interest between the earlier point in time and the latest state as per GHTorrent data. To evaluate the statistical significance of the observed difference, we apply Mann-Whitney U test to both sets of gender diversity index values, as well as Cliff's Delta test.

To answer RQ3, we designed and distributed an online survey to understand motivations and challenges faced when joining and leaving software projects. Our questionnaire comprised of three section of questions. The first section of our survey solicits the motivation of developers to contribute, frequency of participation, reasons for selecting, continue participation, as well as barriers and reasons they have abandoned a software project. To help participants ground their responses, we asked them to answer the above questions for one of the software projects we identified them from. The second section of our survey included questions about how relevant the gender and region of co-contributors is when selecting a project to contribute to. Relating to region, we ask how challenging it is to contributing with people who speak a different language and the usefulness of translation tools to support that interaction. Likewise, we ask for the ease of contributing to projects that have contributors with same gender identity and their advice to encourage women participation in GitHub. In the third section of our survey, we asked demographic questions about their gender identity and the geographic region they contribute to open source from. All questions were optional and presented as either a Likert scale, multiple-choice, or open response question. The survey was designed to be completed in approximately 7 minutes.

We identified survey participants from our GHTorrent sample. Our sample comprised of all contributors from the selected projects for where we can infer the region, gender, and email address in order to contact them. This

9. We use https://github.com/neilernst/cliffsDelta implementation for Cliff's Delta test

subset of contributors was skewed towards some regions (e.g. Northern America was over represented while Micronesia was under-represented). This skew also occurred in distribution of men and women across regions. To gather a representative view spanning multiple regions, we selected 50 men and 50 women from each region. For over-represented groups such as men and Northern America, we randomly identified 50 participants while for under-represented groups (with participants less than 50), we selected all contributors. Overall, we identified 1,562 contributors—of which 1,527 email addresses were valid and did not have an out-of-office reply message.

4 Results

4.1 RQ 1: Gender diversity in different parts of the world

We find that gender diversity of repositories' commit authors are generally low, with global average of diversity index being 0.08 and median of 0. Tables 4 and 5 show the statistical summary of diversity at region levels 1 and 2. Applying Kruskal-Wallis H test on region levels 1 and 2 shows that there is statistically significant difference between regional groups (p=3.89e-115 at region level 1, p=2.60e-61 at region level 2). Table 6 show the adjusted p-values obtained from Mann-Whitney U test with Bonferroni correction for pairs of region level 1. Cliff's Delta test for region level 1 pairs with p$\leq$0.05 yields negligible δ (0.098 for Americas versus Europe, 0.088 for Asia versus Oceania, and 0.132 for Asia versus Europe). Table 7 lists pairs of region level 2 with significant (adjusted p$\leq$0.05) difference as well as their δ. The test results show that while there is statistical difference between regions, including pairs that are close, the difference in Blau index distribution is not substantial except in case of Eastern Europe versus Eastern Asia and Eastern Europe versus South-Eastern Asia.

TABLE 4
Gender diversity index values by region level 1.

Region	Mean	Median	Std. dev.	Min	Max
Africa	0.07	0.00	0.16	0.00	0.50
Americas	0.09	0.00	0.16	0.00	0.50
Asia	0.11	0.00	0.17	0.00	0.50
Europe	0.06	0.00	0.13	0.00	0.50
Oceania	0.08	0.00	0.15	0.00	0.50

Finding: There are statistically significant differences in gender diversity between regions. Although in most cases the difference in Blau index distribution is not substantial, we observe that Eastern Asia and South-Eastern Asia is more diverse than Eastern Europe.

Beyond analysis of gender diversity in specific regions, we are also interested in whether gender diversity correlates with geographic diversity. As the diversity index values for both gender and location are not normally distributed (D'Agostino's K^2 test yields p=0.00 for gender diversity index values as well as region diversity index for all levels of regional grouping), we analyze this by computing Spearman's *rho* between repositories' gender diversity index values and geographic diversity index values at different regional groupings. The result, shown in Table 8, indicates that there is negligible to small negative correlation between gender diversity and geographic diversity. This suggests that different approaches are needed to promote each type of diversity.

TABLE 5
Gender diversity index values by region level 2.

Region	Mean	Median	Std. dev.	Min	Max
Northern Africa	0.07	0.00	0.14	0.00	0.38
Sub-Saharan Africa	0.05	0.00	0.14	0.00	0.50
Northern America	0.09	0.00	0.16	0.00	0.50
Latin America and the Caribbean	0.07	0.00	0.14	0.00	0.50
Western Asia	0.07	0.00	0.15	0.00	0.50
Central Asia	0.04	0.00	0.13	0.00	0.44
Eastern Asia	0.11	0.00	0.17	0.00	0.50
South-eastern Asia	0.11	0.00	0.18	0.00	0.50
Southern Asia	0.09	0.00	0.15	0.00	0.50
Western Europe	0.06	0.00	0.13	0.00	0.50
Northern Europe	0.08	0.00	0.15	0.00	0.50
Eastern Europe	0.05	0.00	0.12	0.00	0.50
Southern Europe	0.08	0.00	0.15	0.00	0.50
Australia and New Zealand	0.06	0.00	0.14	0.00	0.50

TABLE 6
Adjusted p-values for Mann-Whitney U test between regions at level 1. p$\leq$0.05 in bold.

Region	Africa	Americas	Asia	Europe	Oceania
Africa	N.A.	0.755	0.295	1	1
Americas	0.755	N.A.	0.109	**<0.001**	0.340
Asia	0.295	0.109	N.A.	**<0.001**	**0.017**
Europe	1	**<0.001**	**<0.001**	N.A.	0.261
Oceania	1	0.340	**0.017**	0.261	N.A.

Finding: There is no strong correlation between gender and geographic diversity.

4.2 RQ 2: Changes in gender diversity over time

Beyond state of gender diversity based on latest GHTorrent data, we are also interested in how gender diversity changes over time. Figure 2 shows the distribution of gender diversity index score across different repository creation years. We observe no significant change in gender diversity between groups of repositories created between 2008 and 2013, with Kruskal-Wallis H-test yielding a p-value of 0.12. Considering rapid expansion of GitHub in recent years (it has grown from 10 million repositories by end of 2013 to more than 100 million repositories by November 2018 [6]), we decide to focus our analyses of change on the period from 2014 onwards.

To create a baseline for comparison, we use the GHTorrent commit data to identify a set of GitHub users who have authored at least one commit to shortlisted projects by 2014. Since not all shortlisted repositories existed in 2014, this result in a smaller set of 8,338 repositories. We subsequently apply the method used to answer RQ1 to associate each repository to a region and compute gender diversity index values. We note that in some cases a repository may be

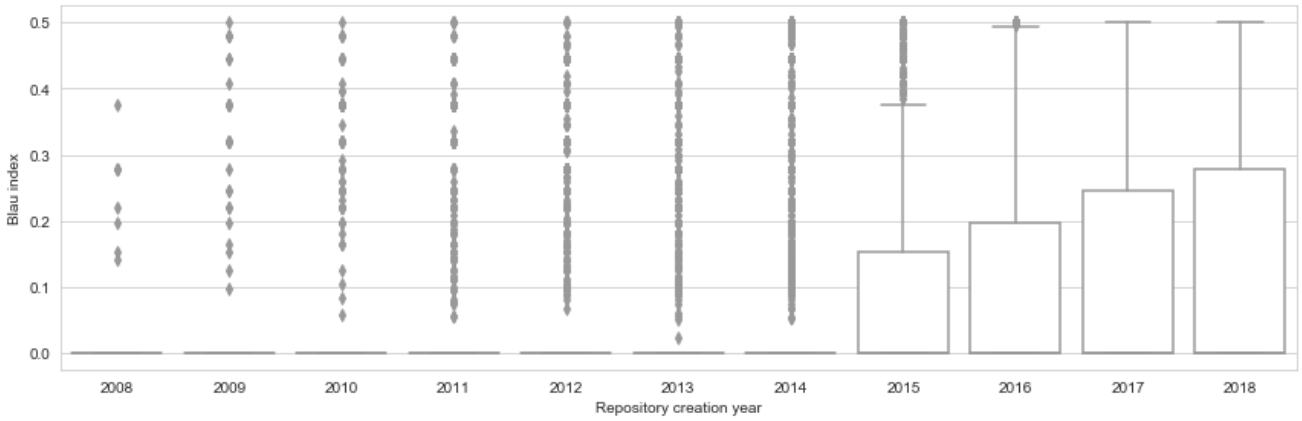

Fig. 2. Gender diversity by repository creation year.

TABLE 7
Pairs of region level 2 with adjusted p-value ≤ 0.05, sorted by δ in descending order. Non-negligible δ are in bold.

Region pair		Adjusted p-value	δ	δ interpretation
Eastern Asia	Eastern Europe	<0.001	0.166	**Small**
South-eastern Asia	Eastern Europe	0.009	0.161	**Small**
Eastern Asia	Western Europe	<0.001	0.145	Negligible
Northern America	Eastern Europe	<0.001	0.141	Negligible
Eastern Asia	Australia and New Zealand	<0.001	0.122	Negligible
Northern America	Western Europe	<0.001	0.118	Negligible
Southern Asia	Eastern Europe	0.005	0.104	Negligible
Eastern Asia	Southern Europe	0.012	0.099	Negligible
Eastern Asia	Northern Europe	<0.001	0.093	Negligible
Northern America	Australia and New Zealand	<0.001	0.093	Negligible
Northern America	Latin America and the Caribbean	0.000	0.084	Negligible
Northern Europe	Eastern Europe	<0.001	0.076	Negligible
Northern America	Southern Europe	0.152	0.067	Negligible
Northern America	Northern Europe	<0.001	0.063	Negligible
Western Europe	Northern Europe	0.000	-0.053	Negligible
Eastern Europe	Southern Europe	0.037	-0.071	Negligible
Latin America and the Caribbean	Eastern Asia	<0.001	-0.112	Negligible

associated with a different region in 2014 and in the latest GHTorrent date if the new commit authors originate from a different location from the old ones. Therefore, we use region information for comparison (e.g. "group of repositories associated with Eastern Europe") instead of comparing changes in individual repositories.

TABLE 8
Spearman's rho between repositories' gender diversity and geographic diversity.

Regional Grouping	rho	p-value
Level 1 (e.g. 'Africa')	-0.06	0.00
Level 2 (e.g. 'Sub-Saharan Africa')	-0.10	0.00
Level 3 (e.g. 'Eastern Africa')	-0.10	0.00
Location (e.g. 'Ethiopia')	-0.11	0.00

Globally, while the mean changes from 0.04 in 2014 to 0.08 at latest GHTorrent date, the median of Blau index of gender diversity remains at 0.00. Mann-Whitney U and Cliff's delta tests yield p-value <0.001 and δ of -0.145, indicating that there's statistically significant but negligible improvement of diversity at global level. To obtain a more detailed breakdown, we also perform analysis at region level 2 and apply Mann-Whitney U test to regions with more than 20 repositories in 2014. The result, shown in Table 9, indicates that while Northern America and South-Eastern Asia shows small but statistically significant change, there is negligible change in other regions. Figures 3 and 4 show a map visualization of the change at region level 2 grouping.

Finding: Since 2014, there has been a small and statistically significant improvement of gender diversity in North America and South-Eastern Asia, but negligible change elsewhere.

Another aspect we are interested in is whether, among commit authors, there is difference in gender balance between older and newer accounts. We investigate this by looking at the account creation years of commit authors whose gender can be resolved, and compute gender composition for each year between 2014-2018 (the latest year for which GHTorrent has complete data). The result, shown in Figure 5, indicates that the percentage of GitHub accounts created by women has remained low throughout the period. This suggests a need to encourage participation of women.

TABLE 9
Changes in gender diversity between 2014 and latest GHTorrent date - region level 2. p and δ calculated only for regions with $\geq$ 20 repositories in 2014. p$\leq$0.05 and non-negligible δ are in bold.

Region Level 2	Repo Count		Median		Mean			p	δ	δ interpretation
	2014	Latest	2014	Latest	2014	Latest	Change			
Northern Africa	6	14	0.00	0.00	0	0.07	0.07	N.A.	N.A.	N.A.
Sub-Saharan Africa	9	35	0.00	0.00	0	0.05	0.05	N.A.	N.A.	N.A.
Latin America and the Caribbean	235	725	0.00	0.00	0.03	0.07	0.04	**<0.001**	-0.121	Negligible
Northern America	4513	11148	0.00	0.00	0.04	0.09	0.05	**<0.001**	**-0.175**	**Small**
Central Asia	1	11	0.00	0.00	0.00	0.04	0.04	N.A.	N.A.	N.A.
Eastern Asia	424	852	0.00	0.00	0.07	0.11	0.04	**<0.001**	-0.124	Negligible
South-eastern Asia	47	86	0.00	0.00	0.05	0.11	0.06	**0.010**	**-0.184**	**Small**
Southern Asia	55	260	0.00	0.00	0.06	0.09	0.03	0.060	-0.100	Negligible
Western Asia	25	75	0.00	0.00	0.02	0.07	0.05	0.058	-0.127	Negligible
Eastern Europe	501	1306	0.00	0.00	0.02	0.05	0.03	**<0.001**	-0.087	Negligible
Northern Europe	821	2134	0.00	0.00	0.04	0.08	0.04	**<0.001**	-0.131	Negligible
Southern Europe	160	493	0.00	0.00	0.03	0.08	0.05	**<0.001**	-0.136	Negligible
Western Europe	1188	3549	0.00	0.00	0.02	0.06	0.04	**<0.001**	-0.114	Negligible
Australia and New Zealand	352	768	0.00	0.00	0.03	0.06	0.03	**<0.001**	-0.122	Negligible
Melanesia	1	0	0	N.A.	0	N.A.	N.A.	N.A.	N.A.	N.A.

Fig. 3. Gender diversity at region level 2 as of 2014. Darker shade indicates higher diversity.

Fig. 4. Gender diversity at region level 2 as per latest data. Darker shade indicates higher diversity.

Finding: Among commit authors with identifiable gender, yearly percentage of account creation by women is low, suggesting that encouragement of participation is still needed.

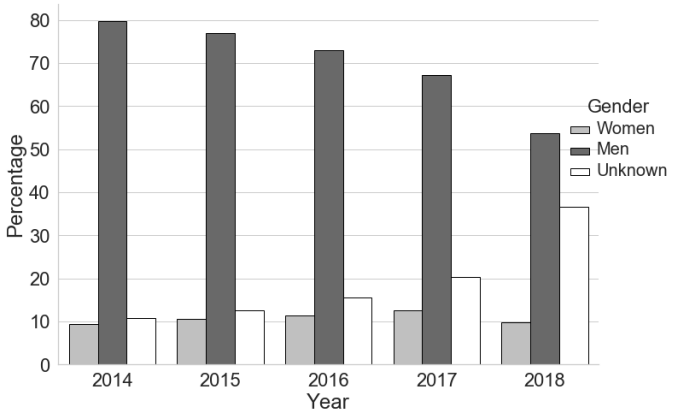

Fig. 5. Gender percentage of commit authors by account creation year, 2014-2018.

4.3 RQ3: Factors influencing participation

To answer RQ3, we gathered all non-empty survey responses associated with a region and/or gender. We processed the data to eliminate aggregates that are few in count for data analysis. We had two types of responses: Likert scale and open-ended. To process Likert scale responses, we transformed ordinal scale to a nominal scale. For example, a 5-point Likert scale of 'Very important, Important, Neutral, Less important, and Not at all important' was converted into Important (combining Very important and Important into one), Neutral, and Not Important (combining Less important and Not at all important into one). The transformed nominal scale was fed as input to Chi-square test to test for statistically significant differences in the responses. All tests were conducted in R and reported at p$<$0.05.

For open response survey questions, three authors conducted a thematic analysis of participant's motivations to contribute, barriers to contribution, and reasons to abandon projects on GitHub. In the first phase, two authors conducted first-cycle descriptive coding on each open-ended response [34]. In the second phase, we performed axial coding to identify core experiences respondents in OSS and the contextual bounds that connects them. In the final phase,

authors discussed codes where responses did not converge by conducting negotiated agreement [35].

We received 122 responses (out of 1,527 emails sent; approximately 8% response rate) during the three weeks in which the survey was active. Our survey garnered one response from woman (total 23) for every four responses from men (total 90). Although we provided the option, no participants in our sample identified their gender as non-binary. Our participants indicate they contribute to open source from around the world including: Europe (46), Asia (29), Americas (21), Africa (12) and Oceania (4), with an overall distribution as shown in Table 11. Some participants preferred not to disclose their gender or geographic region. For our geographic region analysis, we exclude Oceania (due to limited responses). Overall, considering all missing gender and region information, our analysis was limited to 106 responses. In the following subsections, we describe the results from our survey analysis. Quotes from participants are also used to ground our findings.

4.3.1 Project Selection Factors, Motivations, Barriers

We received a range of responses from participants including important factors such as the projects impact, how they are motivated by project alignment, and how they have been inhibited by the community culture.

Project Selection Factors. A majority of developers believe that alignment of project goal to their own is the most important factor for selecting a project. Approximately, 96% of the respondents consider this factor as important while the remaining 4% do not consider it important [χ^2 (1 df) = 86.6, p$<$0.001]. Other factors deemed important are how welcoming the project is (83% important), how easy it is to join the project (81%), and the opportunity to be a part of how software is built (79%).

Although the majority of participants said they did not select a project because they saw it on social media (94% not important) or that their friends or colleagues contribute to that project (67% not important), few acknowledged how other social dynamics did matter. For example, some participants mentioned how important it was to them that a project "supports social equity (P97)" while providing "up-to-date code for others learning (P125)."

Notably, while there were no differences in the reasons to select a project across regions, women developers value selecting a project with friends and colleagues (64% important) more than their men counterparts (75% not important).

Motivations To Contribute. Participants primarily pursued open source software development as their hobby (69 responses), volunteer in the community for free (63), to learn something new (63) or it is their full time job (54) Other less prominent reasons are to get a job (22), meet new people (21), as a part of school or university project (8), and to get paid (6).

From our open responses, participants described their interest in volunteerism as an opportunity to reciprocate what they received from the community in a "socially relevant (P71)" way. One participant goes on to say, "I get so much from the community that I feel where I can I need to give back when I can (P114)."

We identified no difference in participation motivations based on gender, however, regionally some factors stood out. We noted that more participants from Europe (26%) were already employed compared to Africa (8%) [χ^2 (1 df)=12.9, p=0.004]. Similarly, less people from Europe use GitHub to get a job (3%) compared to 11% in Africa [χ^2 (1 df) = 14.0, p=0.002.]. Details on other relations are presented in Table 10.

TABLE 10
Motivation of developers to participate in open source software projects across regions. Each cell reports the percentage of developers motivated by the following factors.

	Europe	Asia	Americas	Africa
my full-time job	26.00	11.00	21.00	8.00
my hobby	21.00	28.00	15.00	19.00
volunteer for free	26.00	20.00	17.00	22.00
learn something new	15.00	24.00	25.00	22.00
school/university project	2.00	1.00	8.00	0.00
help get a job	3.00	8.00	8.00	11.00
meet new people	5.00	6.00	6.00	14.00
get paid	2.00	1.00	0.00	3.00

Barriers to Contribution. From our analysis, we identified 116 barrier statements referring to reasons contributors have decided not participated in some projects or discontinued contributing from others. From these statements we identified 6 themes.

Lack Of Resources. Participants acknowledged that they had limited resources at their disposal to make significant contributions to a project. These resources included time allocation, the lack of project funding, and challenges balancing time spent on projects for a full time job with projects a hobbyist. One participant goes on to describe his work-hobby balance: "I do not do this as a full time job, I just try to commit meaningful changes that helped me in my own projects (P114)." Another describes their funding challenges: "At times I would like to contribute more but it comes down to a lack of funding to put more hours in. (P112)"

Goal Alignment Shift. As contributors grow in their expertise so do their interests and their professional work. For instance, some participants described how there was a predetermined end of their "short-lived project (P26)", but also that they, "have abandoned some open source projects because they have been superseded by other projects or because better options for doing the same thing came along (P13)." Participants did not find useful to stay on a project that was no longer a priority.

Inactivity on Projects. Changing project goals often result in projects being abandoned and eventually becoming inactive. Participants described the signs of dying project: "Decrease in the regularity of contributions from project contributors (P70)." This inactivity on the project went beyond who was contributing. Participants also described significant delay in the code review process from maintainers as a barrier: "In general, having no frequent experienced contributors would make me stop contributing because reviews from experienced developers is one of my main motives to contribute (P118)." Contributors are very interested in contributing to projects a as a learning experience, but when the common experience is, "maintainer just stopped reviewing PRs and abandoned the project (P94)," contributors lose value in participating.

Poor Engineering Environment. Factors related to the engineering environment discouraged contributors. Specifically, participants reported being inhibited by the"complex installation process (P71)", "complex code architecture (P70)", "lack of documentation (P71)", and the "lack of a proper roadmap (P110)." Without proper documentation and a clear roadmap of what the north star of a project is contributors will be misguided like P79 who had a challenge finding the best opportunities to help: "On most [projects I'm] not having a clear understanding of what features would be helpful to work on."

Poor Working Environment. Participants disgruntled by their challenges also recalled the toxic work environments some projects can have: "Sure I have stopped contributing to projects when the maintainers are jerks to me or others. Other thing that have curtailed or stopped me from working on a project are racism, misogynous behavior or unprofessional conduct by maintainers (P43)." A few participants went on to to discuss their 1:1 encounters with project leadership: "The big upstream dependency of this project is maintained by a jerk, so I mostly just maintain the project now, rather than actively add new features (P43)." Although these experiences have been described in low frequency, it is important to note that these experiences can influence how developers decide to contribute like in P43's case.

Unclear Onboarding. The lack of official onboarding documentation processes from maintainers was also discouraging to our participants: "My contribution there was very small, as we did not use it a lot. But I guess this is a good example of the not very well documented project. this is the main obstacle for me when I would like to get involved in some project - not very clear README, missing documentation regarding code discipline for a particular project, not clear rules on how to get involved. That would be for me the main blocker (P98)." When participants reflected on their past experiences with their first project they recalled how challenging it was to join some projects: "The first contact is always the hardest, I mean the totally new newbies always find it intimidating to find and join their first project. (P95)" In short, new contributors to a project have a hard time finding how to get involved.

We did not identify any distinct differences in barriers across region. In descending frequency, our barrier statements came from Europe, Asia, Americas, Africa, and Oceania. Of our participants reporting their contribution region as Asia, the lack of the resource of time was the most prevalent barrier.

Finding: Many of the barriers and motivations for contributing converge across geographic region.

4.3.2 Participation Insights

We identified interesting participation trends from our survey analysis.

Contribution Patterns. Overall, the majority of these respondents contribute to GitHub monthly (79), followed by weekly (22), daily (12) and hourly (4) with no differences in contribution pattern across gender and regions.

TABLE 11
Distribution of survey responses based on gender and region.

Region	Men	Women	Sum
Europe	35	10	45
Asia	25	4	29
Americas	13	7	20
Africa	11	1	12
Oceania	3	0	3
Sum	87	22	109

Continue participation. Once developers have joined a project there are many reasons for developers to continue participation. Developers believe that they continue participating because of interactions with welcoming contributors (91% important) and the global connections they build worldwide (78% important). They further add that having exciting tasks (85% important) as well as low stress level (76% important) in the software project are other reasons to continue participation.

We noticed that developers were quite divided in 'being paid' as an incentive to continue participation. On a closer look we found that being paid is a greater incentive for women (64% find it important) compared to men (65% find it not important). Further, we see that different regions have different reasons to continue participation. Table 12 presents a summary on the relevance of factors in influencing the decision to continue participation. For instance, being paid is not as important for developers from Europe, Asia, and America as it is for Africa. Also, while exciting and challenging tasks are important for all regions, they are more important for developers from Asia and Africa. On a contrary, connecting with people worldwide is not a big motivation for developers from Europe and Americas to continue participation.

TABLE 12
Reasons to continue participation in open source software projects across regions. Each cell reports the percentage of developers that find the following factors important or not important.

	Europe	Asia	Americas	Africa
Interactions with welcoming contributors				
Important	86	96	94	100
Not important	14	4	6	0
Connects with people worldwide				
Important	67	89	77	86
Not important	32	11	23	14
Exciting tasks				
Important	75	100	77	92
Not important	25	0	23	8
Challenging tasks				
Important	84	100	82	100
Not important	16	0	18	0
Being paid				
Important	34	38	21	71
Not important	66	62	79	29

Next, we investigate the importance of shared geographic or gender identity when selecting a software project.

Shared Geographic Region. Overall, having contributors from same geographic region in the project is not important for contribution, albeit subtle differences exist across regions. Having contributors from the same geographic region is least important for Europe, followed by Americas, Asia and

somewhat important for the developers from Africa (see Table 13 for details).

TABLE 13
Relevance of shared/different regional and gender identity - An analysis across geographic regions.

	Europe	Asia	Americas	Africa
Contributors from same geographic region				
Important	9	19	15	40
Not important	91	81	85	60
Working with people who speak a different language				
Challenging	26	50	50	80
Not challenging	74	50	50	20

To understand more on the relevance of same geographic region, we solicited challenges in working with people who speak a different language. We noticed that while overall differences are not discernible, at regional level, the responses are quite divided. Developers from Europe who happen to see no value in having contributors from same region also do not find it challenging working with developers who speak a different language. Developers from Africa, on the other hand, not only find it relatively more important to have developers from the same region, but also have difficulty in interacting with contributors who speak a language different from theirs. Developers in Asia and America are torn amongst themselves in their responses (see Table 13 for details).

To gather a deeper understanding on the subject, we garnered responses on the usefulness of translation tools. Developers hold a mixed opinion on the usefulness of translation tools with no differences across regions. Interestingly, women developers find translation tools more helpful (76%) than men developers (55%).

Shared Gender Identity. We explored the role of same gender identity which is in general seen as not important (91%) but when responses are seen based on gender, 99% men believe that it is not important (only 1% men think it is important) while 37% of women developers believe that it is important. A similar analysis across regions also showed that same gender identity is not at all important for developers from Africa (0%) while it does hold some relevance for other regions: Americas (17%), Europe (11%), and Asia (4%).

Strategies to Support Women. Finally, we wanted to gain a deeper understanding of techniques and mechanisms contributors around the world. Thus, we asked participants about what they think can encourage participation among women on GitHub. We found that some men across regions were very dismissive to this ask saying, "Ask the women. I'm not stopping them (P9)." On the opposition, we also did find some men suggesting how explicit visibility can inspire others, "There were several women highly qualified for any type of project. But if you need any encouragement, perhaps more women will take the initiative to start new open source projects. Maybe it's contagious (P26)." Likewise, we find that most women were interested in women encouraging other women, but through leadership: "More women reviewers. More women acting directly on the governance of large open source projects (P52)."

Finding: For most men, the gender of contributors is not important. Women were more likely to suggest mechanisms that highlight the contributions of women.

5 Design Implications: Tools to Support Inclusion

From our findings, we identify interventions that can better support inclusion across gender and geographic regions.

5.1 Code of Conduct Rewards

Codes of conduct have been encouraged to be used to support a safe environment that can support inclusion [14], [36], [37]. Unfortunately, less than 10% of the top OSS projects actually have one [38]. Participants in our survey also acknowledged that one thing that would encourage inclusion is "Promoting use of and enforcement of code of conduct (P94)." Even fewer projects are transparent about how they enforce these guidelines, if at all.

One approach to enforcing code of conduct usage is rewarding projects that have one. For example, GitHub can offer donation through sponsors program as a reward for projects that have code of conduct. This will provide maintainers with more resources to devote to their role, encourage them to make sure their project is inclusive, and signal to new contributors that a project is safe. Comparatively, this presents a missed opportunity by the projects that have not provided an enforceable code of conduct and thus incentivize those projects to adhere to a new norm.

5.2 Proximity-Based Mentorship

Although in RQ1 we did not find a substantial difference across different geographic regions, survey responses from our participants encourage us to consider what mechanisms can support contributors from specific regions. From the responses, contributors from underrepresented OSS regions are not necessarily resentful. Rather, they would like to empower people from their region to take part in the opportunity to be a builder of software that people around the world use [39], [40]. One participant from Sub-Saharan Africa went as far as to state "Open-source software is a solution for Africa to progress as a continent as quickly as possible while spending less money (P23)".

To support and further activate opportunities such as these, we propose a proximity-based mentorship where mentors and mentees are relatively close in region or even close in cultural dimension (e.g., survival vs. self expression [41]). This experience can take advantage of being in the same shared region by conducting guidance through offline interventions [42]. The duality of fostering both the same community online based on a personal offline experience can further support inclusion.

6 Threats to Validity

Internal Validity. Internal validity of our study can be affected by accuracy of location and gender resolution. We attempt to mitigate this by combining automatic resolution with manual checks, and by using gender resolution tool

that is able to handle names from multiple cultures (e.g. Asian, Eastern European, Arabic). Beyond accuracy, another possible threat to internal validity comes from correctness of the data filled by the GitHub user on their profile.

External Validity. Threats to external validity pertain to the generalizability of result. In our study, we attempt to mitigate this threat by using a sample of repositories that spans a wide variety of software projects. However, it is possible that there are other types of software projects that have different diversity profiles.

Construct Validity. A study design threat we face is the survey instrument we used to collect participant experiences. When we asked participants about the motivations and barriers they faced, we grounded them in reference to the project we identified them making a contribution towards. If their participation was more than a few months prior to the survey, it is likely hard for them to reflect on that experience, especially if it was negative.

7 Conclusion

In this paper, we report findings from our large scale empirical study leveraging quantitative data from GitHub and qualitative data for a targeted survey to developers to report on the gender differences across geographies. Our study finds that there are statistically significant differences in gender diversity between regions. Since 2014, there has been small and statistically significant improvement of gender diversity amongst software contributors in North America and South-Eastern Asia but negligible change elsewhere. We observe that among commit authors with identifiable gender, yearly percentage of account creation by women remains low. A qualitative analysis shows that many of the barriers and motivations for contributing converge across different geographic regions ranging from lack of resources, goal alignment shift to poor working environments and unclear on boarding.

There are two underlying themes we hope this study will achieve. The first is quantifying and setting baseline of current state of GitHub regarding intersection of gender and geography. This will help other researchers build on it and quantify changes in coming years. The second is to create awareness of this problem and hopefully encourage further research by the community towards reducing the gender gap and make software contributions possible by everyone, everywhere. Towards this goal, we are working with people in GitHub and Stack Overflow to help drive some of the concrete observations from our study to alleviate diversity-related issues in the coming years.

8 DATASET AVAILABILITY

In the interest of encouraging others to replicate and build upon our work, we are sharing our data. For now, the data can be found here: https://figshare.com/s/f529853dae2ea4abe601

References

[1] D. Ford, J. Smith, P. J. Guo, and C. Parnin, "Paradise unplugged: Identifying barriers for female participation on stack overflow," in *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, ser. FSE 2016. New York, NY, USA: Association for Computing Machinery, 2016, p. 846–857. [Online]. Available: https://doi.org/10.1145/2950290.2950331

[2] W. DuBow and A. Pruitt, "Ncwit scorecard: The status of women in technology," *NCWIT, Boulder, CO*, 2018.

[3] E. Commission, "Increase in gender gap in the digital sector - study on women in the digital age." http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50224, 2018, study reference: SMART 2016/0025.

[4] P. Raghuram, C. Herman, E. Ruiz-Ben, and G. Sondhi, "Women and it scorecard-india. https://www.nasscom.in/knowledge-center/publications/women-and-it-scorecard-2017.

[5] 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.

[6] GitHub, "Github milestones: A timeline of significant moments in github's history," 2020, retrieved May 28, 2020 from https://github.com/about/milestones.

[7] C. Bird and N. Nagappan, "Who? where? what? examining distributed development in two large open source projects," in *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*. IEEE, 2012, pp. 237–246.

[8] A. Lee, J. C. Carver, and A. Bosu, "Understanding the impressions, motivations, and barriers of one time code contributors to floss projects: A survey," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, May 2017, pp. 187–197.

[9] J. Rigg, *An everyday geography of the global south*. Routledge, 2007.

[10] J. A. Roberts, I.-H. Hann, and S. A. Slaughter, "Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects," *Management science*, vol. 52, no. 7, pp. 984–999, 2006.

[11] I. Steinmacher, T. Conte, M. A. Gerosa, and D. Redmiles, "Social barriers faced by newcomers placing their first contribution in open source software projects," in *Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing*, 2015, pp. 1379–1392.

[12] F. Zlotnick, "Github open source survey 2017," http://opensourcesurvey.org/2017/, Jun. 2017.

[13] C. Mendez, H. S. Padala, Z. Steine-Hanson, C. Hilderbrand, A. Horvath, C. Hill, L. Simpson, N. Patil, A. Sarma, and M. Burnett, "Open source barriers to entry, revisited: A sociotechnical perspective," in *Proceedings of the 40th International Conference on Software Engineering*, 2018, pp. 1004–1015.

[14] D. Ford, R. Milewicz, and A. Serebrenik, "How remote work can foster a more inclusive environment for transgender developers," in *2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE)*. IEEE, 2019, pp. 9–12.

[15] I. Steinmacher, T. U. Conte, C. Treude, and M. A. Gerosa, "Overcoming open source project entry barriers with a portal for newcomers," in *Proceedings of the 38th International Conference on Software Engineering*, 2016, pp. 273–284.

[16] B. Vasilescu, V. Filkov, and A. Serebrenik, "Perceptions of diversity on github: A user survey," in *2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering*. IEEE, 2015, pp. 50–56.

[17] 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 Computer Science*, vol. 3, p. e111, May 2017. [Online]. Available: https://doi.org/10.7717/peerj-cs.111

[18] N. Imtiaz, J. Middleton, J. Chakraborty, N. Robson, G. Bai, and E. Murphy-Hill, "Investigating the effects of gender bias on github," in *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 2019, pp. 700–711.

[19] A. Rastogi, N. Nagappan, G. Gousios, and A. van der Hoek, "Relationship between geographical location and evaluation of developer contributions in github," in *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3239235.3240504

[20] G. Gousios, "The ghtorent dataset and tool suite," in *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE, 2013, pp. 233–236.

[21] M. Ortu, G. Destefanis, S. Counsell, S. Swift, R. Tonelli, and M. Marchesi, "How diverse is your team? investigating gender and nationality diversity in github teams," *Journal of Software Engineering Research and Development*, vol. 5, no. 1, pp. 1–18, 2017.
[22] L. Santamaría and H. Mihaljević, "Comparison and benchmark of name-to-gender inference services," *PeerJ Computer Science*, vol. 4, p. e156, 2018.
[23] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, and D. Damian, "The promises and perils of mining github," in *Proceedings of the 11th working conference on mining software repositories*, 2014, pp. 92–101.
[24] P. M. Blau, *Inequality and heterogeneity: A primitive theory of social structure*. Free Press New York, 1977, vol. 7.
[25] 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.
[26] 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.
[27] R. B. D'agostino, A. Belanger, and R. B. D'Agostino Jr, "A suggestion for using powerful and informative tests of normality," *The American Statistician*, vol. 44, no. 4, pp. 316–321, 1990.
[28] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.
[29] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.
[30] H. Abdi, "Bonferroni and šidák corrections for multiple comparisons," *Encyclopedia of measurement and statistics*, vol. 3, pp. 103–107, 2007.
[31] N. Cliff, *Ordinal methods for behavioral data analysis*. Psychology Press, 2014.
[32] W. G. Hopkins, *A new view of statistics*. Will G. Hopkins, 1997.
[33] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, 2020.
[34] J. Saldaña, *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2009.
[35] J. L. Campbell, C. Quincy, J. Osserman, and O. K. Pedersen, "Coding in-depth semistructured interviews," *Sociological Methods & Research*, vol. 42, no. 3, pp. 294–320, aug 2013.
[36] V. Singh, "Women participation in open source software communities," in *Proceedings of the 13th European Conference on Software Architecture - Volume 2*. New York, NY, USA: Association for Computing Machinery, 2019, p. 94–99. [Online]. Available: https://doi.org/10.1145/3344948.3344968
[37] C. A. Ehmke, "Contributor covenant," 2014, retrieved March 5, 2020 from https://www.contributor-covenant.org/.
[38] V. Singh and W. Brandon, "Open source software community inclusion initiatives to support women participation," in *Open Source Systems*, F. Bordeleau, A. Sillitti, P. Meirelles, and V. Lenarduzzi, Eds. Cham: Springer International Publishing, 2019, pp. 68–79.
[39] Z. Brodsky, "7 chinese open source projects you should know about," 2018, retrieved March 5, 2020 from https://resources.whitesourcesoftware.com/blog-whitesource/7-chinese-open-source-projects-you-should-know-about.
[40] A. Abati, "Made in nigeria," 2017, retrieved March 5, 2020 from https://www.madeinnigeria.dev/.
[41] N. Oliveira, N. Andrade, and K. Reinecke, "Participation differences in q&a sites across countries: Opportunities for cultural adaptation," in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, ser. NordiCHI '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2971485.2971520
[42] Debian, "Debian women mentoring program," 2019, retrieved March 5, 2020 from https://www.debian.org/women/mentoring.

Gede Artha Azriadi Prana is a PhD student in the School of Information Systems, Singapore Management University, under supervision of Associate Professor David Lo. His research focuses on software engineering analytics. Prior to his current study, he worked in the field of software engineering for about a decade.

Denae Ford is a Senior Researcher at Microsoft Research where she works in the Software Analysis and Intelligence (SAINT) team and an Affiliate Assistant Professor at the University of Washington in the Human-Centered Design and Engineering Department. She received her Ph.D. in Computer Science from North Carolina State University. She conducts research in human factors in empirical software engineering including technical interviews, eye-movements in code review, and software developer productivity. Her current research investigates and devises intervention to support inclusive socio-technical interactions in online programming communities.

Ayushi Rastogi is a postdoctoral researcher in the faculty of Electrical Engineering, Mathematics & Computer Science at the Delft University of Technology, the Netherlands. Her research interests include software analytics, empirical software engineering, and mining software repositories. She studies human and social aspects of software development for improving developer productivity and promoting diversity and inclusion.

David Lo is an Associate Professor in the School of Information Systems at the Singapore Management University. He received his Ph.D. in Computer Science from the National University of Singapore. His research interests include software analytics, software maintenance, software testing, social network mining, and cybersecurity.

Rahul Purandare is an Associate Professor in the department of Computer Science and Engineering at the Indraprastha Institute of Information Technology Delhi (IIIT-Delhi). He received his Ph.D. in Computer Science from the University of Nebraska - Lincoln. His research interests include program analysis, software testing, automatic program repair, code search, and code comprehension.

Nachiappan Nagappan is a Partner Researcher at Microsoft Research where he works in the Software Analysis and Intelligence (SAINT) team . He also hold an adjunct faculty appointment at IIIT New Delhi. His research are in the field of Data Analytics and Machine Learning for Software Engineering focusing on Analytics for Empirical Software Engineering, Software Reliability, Software Metrics, and Developer Productivity. His research projects are fairly interdisciplinary and span the spectrum of software analytics ranging from intelligent software design for games; identifying data center reliability to individual developer and team productivity.