Including Everyone, Everywhere

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

Published in:
IEEE Transactions on Software Engineering

DOI:
10.1109/TSE.2021.3092813

Publication date:
2022

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, and Nachiappan Nagappan, *Fellow, IEEE*

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 region. 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 a survey that is strategically targeted to developers in various regions worldwide. Our findings indicate that gender diversity is low across all parts of the world, with no substantial difference across regions. However, there has been statistically significant improvement in diversity worldwide since 2014, with certain regions such as Africa improving at faster pace. 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

The gender gap in the software industry is alarming, garnering attention worldwide. IT companies in India reportedly have women concentration in lower career levels [1]. In the United States, women earning computing degrees rose since the mid-1990s, yet they comprise a quarter of computing professionals [2]. An estimate by the European Commission [3] suggests that if more women enter the digital job market, it could create an annual EUR 16 billion GDP boost for the European economy.

Similar investigations in open source software systems show that despite no significant differences between the work practices of men and women [4] and improved team performance in gender-diverse teams [5], women make up less than 10 percent of core contributors [6]. Further, horizontal and vertical segregation exist [4].

In open source, explorations on gender diversity are all-inclusive, implicitly assuming that the problem remains the same irrespective of the population and project characteristics. However, in this approach, we are likely to miss local achievements in promoting gender diversity and/or problems unique to others. One factor to consider is the geographical region. A study conducted within the European Union shows a disparity in women's participation in digital economies, with Finland and Sweden scoring the highest while Greece and Italy the lowest [3]. This example suggests that digital and online engagement can shift across geographic regions in addition to genders. Thus, inspiring us to ask how this difference in engagement can manifest in open source, specifically.

Our study presents the largest exploration into gender diversity in open source software projects in different parts of the world. We investigate active and collaboratively developed software projects hosted on GitHub to answer:

RQ1: What are the gender and geographic diversity characteristics of open source software projects on GitHub?

The first question is exploratory, presenting the state-of-the-practice on gender diversity and substantiating the need for exploration. Further, we ask questions to open source software contributors to understand:

RQ2: What factors potentially contribute to the differences in gender and geographic-based developer participation?

Our analysis is based on 21,456 carefully selected software projects on GitHub. We use a sequential mixed-methods

- *Gede Artha Azriadi Prana and David Lo are with the School of Computing and Information Systems, Singapore Management University, Singapore, 188065, Singapore. E-mail: arthaprana.2016@phdcs.smu.edu.sg, davidlo@smu.edu.sg.*
- *Denae Ford is with Software Analysis and Intelligence, Microsoft Research, Redmond, WA 98052 USA. E-mail: denae@microsoft.com.*
- *Ayushi Rastogi is with the Faculty of Science and Engineering, University of Groningen, 9712 Groningen, The Netherlands. E-mail: a.rastogi@rug.nl.*
- *Rahul Purandare is with Computer Science and Engineering, Indraprastha Institute of Information Technology, New Delhi, Delhi 110020, India. E-mail: purandare@iiitd.ac.in.*
- *Nachiappan Nagappan is with the Testing, Verification and Measurement Research, Microsoft Research, Redmond, WA 98052 USA. E-mail: nnagappan@acm.org.*

Manuscript received 25 Apr. 2021; revised 20 June 2021; accepted 22 June 2021. Date of publication 28 June 2021; date of current version 19 Sept. 2022. (Corresponding author: Gede Artha Azriadi Prana.) Recommended for acceptance by C. B. Seaman. Digital Object Identifier no. 10.1109/TSE.2021.3092813

approach. First, we quantitatively analyze archived software engineering data of the selected projects to show the state-of-practice of gender diversity worldwide. Next, we survey 1,562 contributors, strategically identified from the selected projects based on gender and geography. We solicit their response in search of factors that can potentially contribute to the differences in developer participation based on gender and geography worldwide.

Our analyses of a decade of development activities on GitHub show small but significant improvements in gender diversity in the last five years. While we celebrate the positive change, it is important to remember that we are far from reaching gender balance. Our study further shows that gender diversity changes over time have not been the same across regions. Some regions such as Eastern Asia and Northern America are (relatively) ahead in gender diversity, while others such as Eastern Europe and Sub-Saharan Africa are still catching up. These differences are also reflected in our investigation of gender and regional related motivations and challenges.

This comprehensive guide of gender-geographic challenges and opportunities can direct future in-depth explorations catering to sub-population needs. For example, one of the opportunity identified here is having a code of conduct. Having a code of conduct can support a two-pronged approach of: 1) allowing lurkers interested in contributing (e.g., including women and other marginalized developers) to feel more comfortable in contributing since they know there are guidelines that can protect them from toxic interactions and 2) signal to developers who are already in the community (e.g., including those that may have been inciting toxic interactions) that there will be repercussions for their actions. Solutions such as these can have a long-term impact to minimize gender gap and uplifting society.

Our contributions are as follows:

(1) We present an analysis of the activity and experiences at the intersection of gender and global geographic region.
(2) Large-scale global analysis of regional gender diversity spanning 21,456 active GitHub repositories and 70,621 commit authors.
(3) Global survey of factors that contribute to the differences in gender and geographic-based developer participation, with 122 respondents across 5 large geographic regions and across genders.
(4) A discussion of actionable implications of how to support OSS sub-communities across gender and geographic regions.
(5) A publicly available dataset to encourage further investigations.

2 Background and Related Work

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 [7], barriers to participation [8], and how developers contribute to open source [9]. 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 motivation and 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 important studies that have shaped this area of research.

2.1 Motivation to Contribute to Open Source Software

Motivation in software engineering has been subject to numerous studies, including several systematic reviews [10], [11], [12]. The existing body of works include a number of studies focusing specifically on motivation of OSS contributors. For example, in a 2002 study, Hars and Ou [13] surveyed open source developers and found that their motivation for contributing are diverse – while students and hobbyists tend to be internally motivated, there are also a large number of developers who are motivated by external rewards. Lakhani and Wolf [14] surveyed 684 OSS developers and found that the strongest type of motivation among the respondents are enjoyment-based intrinsic motivation. prominence of enjoyment-based intrinsic motivation. Von Krogh *et al.* [15] examined prior literature on OSS developers' motivation to contribute, and proposed 10 clusters of motivation types categorized into intrinsic motivation, internalized extrinsic motivation, and extrinsic motivation. Barcomb *et al.* [16] surveyed episodic (non-habitual) OSS volunteers and found that intention to remain are positively associated with social norms, satisfaction, and community commitment. Further, they also found some differences based on participants' gender. Most recently, a study by Gerosa *et al.* [17] investigated how main motivations of OSS contributors as a group change over the years and how OSS contributors' individual motivations change as they become more experienced. They found that among OSS contributors, some motivations related to social aspect has gained popularity in recent years. They also found that experienced OSS contributors tend to be motivated by intrinsic factors such as altruism, unlike new contributors who tend to place higher importance on factors such as career and learning. These studies facilitate better understanding of what drives people to contribute to OSS projects, what approaches project owners can take to attract contributors, and how these contributors can be retained.

2.2 Barriers to Participation in Open Source

A number of studies investigate barriers that can prevent developers from participating to open source. These barriers have been identified in tools, processes [18], and social collaborations [8]. For example, a study by Terrell *et al.* [19] found that while women have higher overall acceptance rate of pull requests, their acceptance rate is lower than men when their gender are identifiable and they are not insiders to a project. Another study by Rastogi *et al.* [20], which analyzes pull requests from 17 countries, found that acceptance rate of contributions can vary significantly depending on the contributor's country of origin, and are higher when when they are evaluated by developers from the same country. The study however does not analyze gender as a factor, as they noted that including only pull requests for which gender data can be obtained will result in sample size that

is too small. Other studies examine barriers such as those affecting acceptance of contribution from newcomers [8] or those affecting underrepresented communities [21]). These studies not only help in raising awareness of existence of such barriers, but they also help in identifying the source of problem. Further, studies such as [22] also propose solutions that can be adopted by OSS community to mitigate such barriers.

2.3 Diversity in Open Source Software Projects

In line with increasing awareness regarding the importance of diversity in broader work context, diversity in open source software projects has gained increasingly widespread attention. Starting from the awareness of diversity and particularly the demographic attributes of developers [23], [24], today improving diversity is seen as a goal for fairness [19] as well as improved productivity [25]. Many studies relating to gender diversity and the lack thereof followed, discussing its relevance [19], state of diversity among popular OSS projects [6], male and female OSS contributors' perceptions of other contributors [26], perceptions of women core developers in OSS projects [4], and the impediments to improve gender diversity [27].

All these studies identify challenges and needs of underrepresented communities. We conduct a comparison outlining the distinction between our work and closely-related prior work in Table 16 in Appendix A.

Our study has common elements to the developer survey on Stack Overflow users [9] but their findings are not synonym to open source. That said, the report does not provide empirical data to support the full scope of motivations and how they persist across genders or regions. Our work provides novelty by conducting an analysis of the activity and experiences at the intersection of gender and global geographic region. 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

We used a convergent mixed-methods study approach to answer our research questions [28]. We identified active OSS projects that are likely to be non-toy projects, resolved the gender as well as location of the project contributors, and then distributed a survey to understand their motivations and challenges. The following subsections describe each of these steps in detail.

3.1 Identification of Suitable GitHub Repositories

3.1.1 Initial Set of Repositories

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], [19], [25]). 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). This is to reduce probability that the project is a "toy" repository (e.g., a user trying a programming tutorial) or a student programming assignment (which usually lasts less than a semester).
- The repository has at least one commit from the beginning of 2018 or later. This is to reduce probability that the project is inactive.
- The repository has at least 10 commits from 4 or more distinct commit authors, none of which are marked 'fake' or 'deleted' GHTorrent.
- The repository is not a fork. We chose not to evaluate forks since we are interested in "core" contributors of a project. In addition, contributions to forks are not always integrated back to the original project and there may also be redundant development between forks and original projects [29].

The above criteria were set to reduce probability of including "toy" projects while avoiding potential elimination of active non-toy projects. Considering rapid growth of GitHub in recent years, we believe the criteria still allows newer OSS projects, for example those created in 2018, to be included in the study.

3.1.2 Location Resolution of Commit Authors

We subsequently attempt to resolve the location 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 other fields in the following order:

(1) *location* field. For example, if the commit author specifies "Seattle" as their location, the country assigned will be USA. If they specify "Tokyo", the country assigned will be "Japan".
(2) Latitude and longitude (*lat* and *long* fields in GHTorrent data, respectively).
(3) *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.
(4) *email* field. For example, if the author's email address uses an Australian government domain, the country assigned will be Australia.

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 and Associated Commit Authors

Shortlisted Repositories				
	Min	Max	Mean	Median
No. of Commit Authors	4	109	6.16	5
No. of Commits	22	301692	363.27	170
Creation year	2008	2018	2014.63	2015
Commit Authors of Shortlisted Repositories				
Total commit authors count				70,621
Commit authors with resolvable location				58,498
Commit authors with resolvable gender				65,132
Commit authors with resolvable gender and location				56,866

Considering differences in culture and other factors that may exist within a region (for example, North American countries versus Latin American countries, Western European countries versus Eastern European countries), we also assign three levels of region information to each commit author based on the taxonomy of regions specified by United Nations Statistics Division.[1] 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). Our intention is to facilitate analyses at finer granularity instead of treating a continent (e.g., America, Asia, Europe) as a unit.

3.1.3 Gender Resolution of Commit Authors

For the commit authors' gender, resolution is attempted by identifying first name portion of the commit author's name. This is followed by resolution of gender using *genderize.io*,[2] which has been reported to have high accuracy [30], [31] and has been used in various studies related to gender representation (e.g. [32], [33], [34]) as well as in the media[3] 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 using a combination of CC-CEDICT[4] (for Chinese characters) and Google Translate.[5]

As an additional measure to evaluate *genderize.io*'s accuracy, one of the authors randomly selected five sample repositories for manual validation. The repositories are associated with a total of 57 contributors from different regions (15 from Americas, 12 from Asia, 18 from Europe, 4 from Oceania, and 8 with unknown region). Each repository is 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 overall the manual analysis results match *genderize.io*'s results 89.5 percent of the time, with 100 percent match on European and Oceanian contributors, 91.7 percent on Asian contributors, 80 percent on contributors from Americas. In case of contributors whose location is unresolvable, there is 75 percent agreement between manual resolution and *genderize.io*'s prediction based on contributors' names.

3.1.4 Final Selection of Repositories

Following this, we apply further filtering for repositories for which both gender and location can be resolved for at least 75 percent of the commit authors. Considering that not all repositories on GitHub are software project repositories [35], 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.

3.1.5 Calculating Gender Diversity of Commit Authors

To measure the gender diversity of commit authors from different regions, we use the Blau diversity index [36] which has also been used in several works in software engineering domain [25], [37], [38]. In simple terms, the index specifies the probability that two randomly-selected members of a group would belong to different categories. It is defined as $1 - \sum_{i \in \{m,f\}} p_i^2$, where p_i^2 are proportion of men and women ("M" and "F", respectively) among commit authors.

During calculation, we disregard unknown values. For example, if a region 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 set of commit authors from a region comprise two men, two women, and one person with unidentified gender, the gender diversity index will be 0.5, which is the maximum value.

To check whether the diversity of commit authors is independent from region, we apply the Chi-squared test to analyze distribution of the two genders across regions, and subsequently computed Cramér's V [39] to measure association strength between gender and region at both region levels. For this analysis, we include commit authors whose location and gender are resolvable (56,866 commit authors

1. https://unstats.un.org/unsd/methodology/m49/
2. http://www.genderize.io
3. https://genderize.io/use-cases
4. https://cc-cedict.org/wiki/
5. https://translate.google.com/

TABLE 3
Commit Author Region and Gender in Shortlisted Repositories, Sorted by Region Level 1

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

comprising 53,426 men and 3,440 women). Exclusion of commit authors with unknown gender is done for consistency with Blau diversity index computation, while exclusion of commit authors with unknown location is done since we are interested in variation between regions worldwide.

Since we note that most projects (70.27 percent) have a majority region at region level 1, i.e. level 1 region from which more than half commit authors originate, we also performed a repository-oriented diversity analysis to provide additional perspective. To do this, 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 {"Europe", "Americas", "Americas", "Americas", "Asia"}, then the repository will be associated with Americas. Afterwards, we compute the diversity index of each repository. To test statistical significance and effect size of the difference, we first apply Kruskal-Wallis H test on groups of repositories associated with each level 1 regions. We subsequently applied Mann-Whitney U test [40] with Bonferroni correction [41] to compare different pairs of region level 1. Afterwards, we computed Cliff's Delta [42] on level 1 region pairs[6] with statistically significant difference to discover the effect size.

After we conducted our initial analysis on the commit authors, we also considered following the line of research of Trinkenreich *et al.* [43] by investigating activities of non-technical contributors. We extracted data of GitHub users who had never authored a commit to the shortlisted sample repositories but had created, changed, or commented on issues and merged pull requests associated with the sample repositories. We excluded user IDs that are marked "fake" and "deleted" in GHTorrent. We found 299,159 users that are not also commit authors. Out of this group, 30.59 percent has both unresolvable gender and location. Beyond this, 21.56 percent has unresolvable location although their genders are resolvable, and 9.54 percent has unresolvable gender although their locations are resolvable. Table 4 shows the breakdown of this non-author group by region level 1, along with the Blau index of the users in this group whose gender is resolvable. We note that for members of this group with resolvable gender and location, the vast majority is male, and like the case with commit authors, there is low diversity in the various regions studied. However, due to the large percentage of users with unknown gender and/or location among this group, we decided not to analyze this group and to focus our analysis solely on commit authors.

6. We use https://github.com/neilernst/cliffsDeltaimplementation for Cliff's Delta test.

TABLE 4
Diversity and Counts Of Contributors Other Than Commit Authors by Region Level 1

Region Level 1	Count				%	Blau index
	M	W	Unknown	Total		
Europe	43873	1402	10303	55578	18.6	0.06
Oceania	3359	121	1022	4502	1.5	0.07
Americas	44859	2430	8909	56198	18.8	0.10
Africa	1432	87	387	1906	0.6	0.11
Asia	15424	1351	7860	24635	8.3	0.15
Unknown	59486	4857	91428	155771	52.2	0.14

Entries are ordered by non-decreasing Blau index value. Blau index of 0.5 indicate maximum diversity (50 percent men, 50 percent women).

3.1.6 Examining Correlation Between Geographic and Gender Diversity

We are also interested in whether a repository's gender diversity correlates with its geographic diversity. As the Blau index values of repositories' contributor gender and location diversity are not normally distributed (D'Agostino's K^2 test [44] 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 rank correlation test [45] between repositories' gender diversity index values and geographic diversity index values at different regional groupings. We use *SciPy* [46] implementation of these statistical tests, and follow scale of interpretation of ρ used by Camilo *et al.* [47] ($\pm$ 0.00-0.30: Negligible, $\pm$0.30-0.50: Low, $\pm$0.50-0.70: Moderate, $\pm$0.70-0.90: High, and $\pm$0.90-1.00: Very high).

3.1.7 Examining Gender Diversity Changes over Time

Beyond state of gender diversity based on latest GHTorrent data, we are also interested in how gender diversity changes over time. 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 [48]), 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. We subsequently apply the same approach used for RQ1 to compute diversity index values for different regions in 2014. We then perform Kruskal-Wallis H test to evaluate the

TABLE 5
Distribution of Surveyed Commit Authors at Region level 2

Region Level 1	Region Level 2	M	W	Unknown
Africa	Northern Africa	50	3	1
Africa	Sub-Saharan Africa	50	2	2
Americas	Latin America and the Caribbean	50	50	17
Americas	Northern America	50	50	50
Americas	Others	3	0	0
Asia	Central Asia	22	0	1
Asia	Eastern Asia	50	50	50
Asia	South-eastern Asia	50	30	9
Asia	Southern Asia	50	50	15
Asia	Western Asia	50	11	5
Europe	Eastern Europe	50	49	20
Europe	Northern Europe	50	50	20
Europe	Southern Europe	50	39	7
Europe	Western Europe	50	50	50
Oceania	Australia and New Zealand	50	39	10
Oceania	Melanesia	3	0	0
Oceania	Polynesia	4	0	0
Unknown	Unknown	50	50	50

statistical significance of the difference in diversity between 2014 and latest state. Afterwards, we calculate the effect size using Cliff's Delta.

3.1.8 Examining Gender Diversity of Older versus Newer Accounts

An additional 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 all commit authors of the shortlisted projects, and compute gender composition for each year between 2014-2018 (the latest year for which GHTorrent has complete data).

3.2 Globally-Distributed Developer Survey

3.2.1 Protocol

To understand motivations and challenges faced by developers of different genders in various regions when joining and leaving software projects, we designed and distributed an online survey. The survey comprised three section of questions. The first section solicits the motivation of developers to contribute, frequency of participation, reasons for selecting a particular project, continue participation, as well as barriers and reasons they have abandoned a software project. We build upon previous surveys on barriers and experiences in online programming communities to develop our survey questions in this section [9], [49], [50]. 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. This section of questions is inspired by how peer parity can encourage participation of people from a shared background or identity [51]. Relating to region, we ask how challenging it is to contribute with people who speak a different language and the usefulness of translation tools to support that interaction. Likewise, we asked about the ease of contributing to projects that have contributors with same gender identity and their advice to encourage women participation in GitHub. Finally, in this section we asked all respondents about what should be done to encourage more women in OSS which is aligned with previous surveys [4], [50]. In asking all respondents, we understand better how to approach interventions that not only serve women, but also those of other marginalized identities across geographic regions. 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.

TABLE 6
Distribution of Surveyed Commit Authors at Region Level 1

Region Level 1	M	W	Unknown
Africa	100	5	3
Americas	103	100	67
Asia	222	141	80
Europe	200	188	97
Oceania	57	39	10
Unknown	50	50	50

3.2.2 Participants

We identified survey participants from our GHTorrent sample. Our sample comprised all contributors from the selected projects for whom we can infer region, gender, and email address to contact them. The distribution of contributors was skewed towards some regions (e.g., Northern America was over-represented while Micronesia was underrepresented). We observed this skew also in the distribution of men and women across regions.

To gather a representative sample 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 underrepresented 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. The distribution at region level 2 is shown in Table 5, while the total for each region level 1 is shown in Table 6.

We received 120 responses (out of 1,527 emails sent; approximately 8 percent response rate) in three weeks. On reviewing the responses, two authors manually analyzed half of the survey responses each for anti-patterns (e.g., all responses are empty or have the same value for all questions). We found two responses with all empty values which we discarded from analysis. We did not observe any other patterns in survey responses. We used 118 responses after discarding the two empty responses.

Our survey garnered approximately one response from a woman (total: 23) for every four responses from men (total: 90). Although provided with an option, no participants in our sample identified their gender as non-binary. Our

TABLE 7
Distribution of Survey Responses Based on Gender and Region

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

TABLE 8
Gender Diversity (or Blau) Index Arranged in Non-Decreasing Order by Region (Level 1)

Region	Blau index	Commit authors (% distribution)
Africa	0.08	364 (1%)
Europe	0.08	24350 (34%)
Oceania	0.10	1880 (3%)
Americas	0.14	26607 (38%)
Asia	0.15	5297 (7%)
Unknown	0.17	12123 (17%)

Blau index of 0.5 indicate maximum diversity (50 percent men, 50 percent women).

participants have contributed to open source from around the world, including Europe (46), Asia (29), Americas (21), Africa (12), and Oceania (4), with an overall distribution shown in Table 7. Some participants preferred not to disclose either gender or geographic region; hence the total count in Table 7 is lower than the number of responses received.

3.2.3 Analysis

We had two types of responses: Likert scale and open-ended. To process Likert scale responses, we transformed an ordinal scale into 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). This way it is easier to (statistically) distinguish factors deemed important from not important, in addition to the overall distribution. Similarly, other Likert scale questions were processed.

The transformed nominal scale was fed as input to the Chi-square test to test statistically significant differences in the responses. All tests were conducted in R and reported at $p < 0.05$. For data analysis, we analyze aggregates for which we can draw meaningful inferences. Since gendered responses from Oceania are fewer in the count, we remove them from statistical analysis.

For open response survey questions, the 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, four authors independently conducted first-cycle descriptive coding [52] (i.e., summarizing the topic of each response as code) on each open-ended response. In the second phase, one author performed axial coding (i.e., relating the codes to each other) to connect core experiences respondents had in OSS. In the final phase, three authors discussed codes where responses did not converge by negotiation [53].

TABLE 9
Gender Diversity Index Values Arranged in Non-Decreasing Order by Region (Level 2)

Region Level 1	Region Level 2	Blau Index	Commit authors
Americas	Others	0.00	5
Oceania	Melanesia	0.00	5
Oceania	Polynesia	0.00	5
Asia	Central Asia	0.06	34
Europe	Eastern Europe	0.06	3858
Europe	Southern Europe	0.06	2314
Africa	Sub-Saharan Africa	0.07	273
Asia	Western Asia	0.07	529
Europe	Western Europe	0.08	10637
Americas	Latin America and the Caribbean	0.09	2547
Europe	Northern Europe	0.10	7541
Oceania	Australia and New Zealand	0.10	1870
Africa	Northern Africa	0.11	91
Asia	Southern Asia	0.11	1463
Asia	South-eastern Asia	0.13	686
Americas	Northern America	0.14	24055
Asia	Eastern Asia	0.20	2585
Unknown	Unknown	0.17	12123

Blau index of 0.5 indicate maximum diversity (50 percent men, 50 percent women).

4 Results

4.1 RQ1: What are the Gender and Geographic Diversity Characteristics of OSS Projects on GitHub?

4.1.1 Regional Variations

We find that gender diversity of repositories' commit authors are generally low worldwide, as shown in Tables 8 and 9. Through Chi-squared test, we found relationship between gender and region ($p = 6.25e\text{-}56$ at region level 1 and $p = 1.30e\text{-}78$ at region level 2) but negligible association strength (Cramér's V result of 0.07 at region level 1 and 0.08 at region level 2).

The result of our repository-oriented additional analysis at region level 1, shown in Table 10, demonstrates similar ordering from least to most diverse regions. We find that this approach produce overall result that is consistent with result of our previous, region-oriented approach. There is statistically significant difference among regions overall,

TABLE 10
Gender Diversity Index Values by Region Level 1, Computed by Associating Project With Most Frequent Contributor Location

Region	Mean	Median	Std. dev.	Min	Max
Europe	0.06	0.00	0.13	0.00	0.50
Africa	0.07	0.00	0.16	0.00	0.50
Oceania	0.08	0.00	0.15	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

TABLE 11
Spearman's ρ Between Repositories' Gender Diversity and Geographic Diversity

Regional Grouping	ρ	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*

** indicates p-value < 0.001.*

TABLE 12
Changes in Gender Diversity of Commit Authors Between 2014 and Latest GHTorrent Date - Region Level 2

Region Level 2	Diversity Index			Users	
	2014	Latest	Change	2014	Latest
Northern Africa	0.00	0.11	0.11	9	91
Sub-Saharan Africa	0.00	0.07	0.07	55	273
Latin America and the Caribbean	0.04	0.09	0.05	563	2547
Northern America	0.09	0.14	0.05	7250	24055
Americas (Others)	N.A.	0.00	N.A.	0	5
Central Asia	0.00	0.06	0.06	6	34
Eastern Asia	0.18	0.20	0.02	772	2585
South-eastern Asia	0.12	0.13	0.01	159	686
Southern Asia	0.08	0.11	0.03	207	1463
Western Asia	0.02	0.07	0.05	113	529
Eastern Europe	0.03	0.06	0.03	962	3658
Northern Europe	0.08	0.10	0.02	2128	7541
Southern Europe	0.05	0.06	0.01	562	2314
Western Europe	0.05	0.08	0.03	2963	10637
Australia and New Zealand	0.08	0.10	0.02	573	1870
Melanesia	0	0.00	0.00	2	5
Polynesia	N.A.	0.00	N.A.	0	5
Unknown	0.11	0.17	0.06	2439	12123

N.A. indicates regions for which Blau index cannot be computed since there are no users at the time.

($p = 3.89e\text{-}115$ in Kruskal Wallis H test). We found three pairs with statistically significant difference in Mann-Whitney U test (Americas versus Europe, Asia versus Oceania, and Asia versus Europe, all of which have $p < 0.001$). However, we observe negligible effect sizes on Cliff's Delta test (δ of 0.098 for Americas versus Europe, 0.088 for Asia versus Oceania, and 0.132 for Asia versus Europe).

Finding: Gender diversity is low worldwide, and while there is apparent difference in diversity across regions (with Asia and Americas being highest), statistically the difference is not substantial.

4.1.2 Correlation between Geographic and Gender Diversity

The result of our analysis of correlation between geographic and gender diversity, shown in Table 11, shows negligible to small negative correlation between gender diversity and geographic diversity. This suggests that project teams that accept contributors from different regions may still be homogeneous in terms of gender, and vice versa, indicating that different approaches are needed to promote each type of diversity.

Fig. 1. Gender diversity at region level 2 as of 2014. Darker shade indicates higher diversity.

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

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

4.1.3 Gender Diversity Changes Over Time

Table 12 shows the change in Blau index at region level 2, while Figs. 1 and 2 show the map visualization. We note that there is general trend of improvement, with most regions showing increase in Blau index value, and none show a decrease. We found that the difference between 2014 Blau index values of the various regions and the latest values is statistically significant ($p = 0.03$), and Cliff's Delta calculation indicate large effect size ($\delta = 0.47$). However, as shown in Table 12, in terms of absolute value, there is still much room for improvement; most regions see an increase in Blau index values of less than 0.10 since 2014, with the exception of Northern Africa, which improved by 0.11.

Finding: Globally, the increase in gender diversity in OSS projects is statistically significant with large effect size, however there is still much room for improvement.

4.1.4 Gender Diversity of Older versus Newer Accounts

Fig. 3 shows the breakdown of commit author accounts by creation year and gender. The percentages indicate that the number of GitHub accounts created by women has

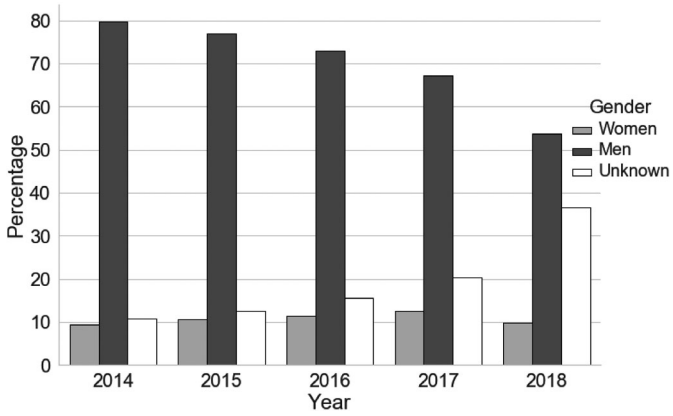

Fig. 3. Gender percentage of commit authors by account creation year, 2014-2018.

remained low throughout the period. This suggests a need to encourage participation of women.

Finding: Among commit authors with identifiable gender, yearly percentage of account creation by women is around 10 percent, suggesting that encouragement of participation is still needed.

4.2 RQ 2: What Factors Potentially Contribute to The Differences in Geographic- and Gender-based Developer Participation?

We received a range of survey responses from participants that include important factors such as the projects impact, how they are motivated by project alignment, and how they have been inhibited by the community culture. In this section we report the results of our analysis which was done at two levels: globally and regionally. Our objective is to obtain both a global view of factors affecting developer participation, as well as view of any region-specific characteristics that can be utilized to promote participation from particular regions.

4.2.1 Global Findings

Overall, we find that the majority of survey 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.

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 percent of the respondents consider this factor as important while the remaining 4 percent 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 percent important), how easy it is to join the project (81 percent), and the opportunity to be a part of how software is built (79 percent).

Although the majority of participants said they did not select a project because they saw it on social media (94 percent not important) or that their friends or colleagues contribute to that project (67 percent 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)."

Finding: To encourage participation in a project, goal alignment and creation of welcoming community will be more effective than promotion in social media.

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)."

Motivations to Continue Participation. Once developers have joined a project there are many reasons for developers to continue participation. The factor that is considered most important is interactions with welcoming contributors (91 percent of participants consider this important). This is followed by availability of exciting tasks (considered important by 85 percent) and the global connections they build worldwide (78 percent important). Low stress level (considered important by 76 percent) is another common consideration to continue participation.

Finding: While developers may participate in a project for variety of reasons, ensuring continued participation requires project owners to maintain welcoming community, ensuring availability of exciting tasks, and minimizing stress to contributors.

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.

4.2.2 Gender and Regional Related Motivations and Challenges

We found that women developers place high importance on social aspects related to OSS projects as an aspect to consider before participating. Women value selecting a project with friends and colleagues more than men (64 percent of women participants consider this important, compared to only 25 percent of men). Beyond this, 37 percent of women developers believe that shared gender identity with fellow contributors as important, while only 1 percent of men consider it important. Analysis across regions showed that same gender identity is not at all important for developers from Africa (0 percent) while it does hold some relevance for other regions: Americas (17 percent), Europe (11 percent), and Asia (4 percent). Beyond the social aspect, we also found that being paid is a greater incentive for women (64 percent find it important) compared to men (35 percent find it important).

We also 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 question 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)." Additional details about this finding can be found in Appendix B, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/10.1109/TSE.2021.3092813.

Finding: Shared gender identity, working with friends and colleagues, and being paid is more important for women than men.

4.2.3 Regional Variation in Motivations and Challenges

Motivation to Participate in OSS Projects. Table 13 shows the developers' motivation to participate in OSS projects, broken down by region. We find that motivation to contribute to OSS as a full-time job is less common outside of Europe and the Americas. In addition, developers from Africa placed relatively higher importance on networking (i.e., "meeting new people") compared to developers from other regions.

Motivation to Continue Participation in OSS Projects. Table 14 shows the developers' motivation to continue their participation in OSS projects, broken down by region. We note that there are regional variations regarding importance of various factors. For example, while exciting and challenging tasks are important for all regions, they are more important for developers from Asia and Africa. On the other hand, connecting with people worldwide is not a big motivation for developers from Europe and Americas to continue participation.

We also found regional differences between what motivates developers to *participate* and what motivates developers to *continue* participation. This difference is in line with

TABLE 13
Motivation of Developers to Participate in Open Source Software Projects Across Regions

	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

Each cell reports the percentage of developers motivated by the following factors.

TABLE 14
Reasons to Continue Participation in Open Source Software Projects Across Regions

	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

Each cell reports the percentage of developers that find the following factors important or not important.

Gerosa *et al.*'s finding [17] regarding shift in motivation of OSS contributors as these contributors gain tenure. For instance, as shown in Table 13, the percentage of African developers who participate in OSS as full-time job, to help get a job, or to get paid is relatively small. However, Table 14 shows that being paid is an important consideration for African developers to continue participation, much more so than it is for developers from Europe, Asia, and America. This suggests that while African developers may start participating in OSS projects as a hobby, to volunteer, or to learn something new, monetary rewards are important to maintain long-term participation. As another example, while a small percentage of Asian developers stated "meeting new people" as a reason to participate in OSS projects, 89% reported connecting with people worldwide as a reason to continue participation — a percentage similar to developers in Africa (86%).

Finding: Some form of funding for participation in OSS projects can be particularly effective to promote continued participation of developers from Africa.

TABLE 15
Relevance of Shared Regional Identity and Language 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

Relevance of Shared Regional and Linguistic Identity. 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 15 for details).

We also solicited challenges in working with people who speak a different language, and 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 fellow developers from the same region in the project, but also have difficulty in interacting with contributors who speak a language different from theirs. Meanwhile, developers in Asia and America are evenly split in their responses (see Table 15 for details). We also found that developers overall hold mixed opinion on the usefulness of translation tools, with no differences across regions. However, there is a difference across genders. We found that 76 percent of women developers find translation tools helpful, but only 55 percent of men developers do so.

Finding: Provision of better translation tools will be particularly helpful to encourage participation of women developers worldwide, as well as participation of developers from Africa.

5 Discussion

5.1 Summary of Findings

Our result for RQ1 did not show substantial difference across different geographic regions. We note that the set of commit authors with unresolved location has higher apparent Blau index compared to sets from known regions. A factor that contributes to this is the high percentage of users in the set whose gender is also unresolved (31.82 percent, as shown in Table 3). Since the Blau index calculation ignores "Unknown" gender, and majority of commit authors are probably men (based on proportions of commit authors whose gender and location can be resolved), we believe the high percentage of unknowns increases apparent women-to-men ratio in favor of women. This subsequently increases the Blau index of the group with unknown location.

As for the observed diversity improvement during the period analyzed in this work, we believe it is influenced by a combination of factors. First, in recent years there has been increasing interest in promotion of diversity in computing. This includes efforts by non-profit organizations (such as Girls Who Code[7], Women Who Code,[8] NCWIT,[9] and ACM-Women[10]), programs targeted at school students [54], [55], initiatives by universities to improve diversity in their own programs [56], [57], [58], as well as efforts by various organizations worldwide to hire more diverse staff. This occurs along the growth of the software industry including in previously underrepresented regions such as Africa [59], with GitHub itself seeing a drastic increase in popularity outside the United States[11]. These factors help attract more diverse talents into computing, including women from underrepresented regions. Nevertheless, as the data shows, there is still much room for improvement.

Related to RQ2, survey responses from our participants encourage us to consider what mechanisms can support contributors from specific regions. In summary, our findings highlight three approaches that should be utilized to better support inclusion across gender and geographic regions. They are:

(1) Development of friendlier communities, especially towards newcomers.
(2) Highlighting of role models from marginalized communities.
(3) Augmentation of existing automated software engineering techniques to incorporate social factors.

5.2 Opportunities Ahead

5.2.1 Development of Friendlier Communities

There are several ways to encourage development of friendlier, more welcoming communities. Creation and enforcement of codes of conduct are an example of a way to promote a safe environment that can support inclusion [21], [60], [61]. Having a code of conduct can support a two-pronged approach of: 1) allowing lurkers interested in contributing (e.g., including women and other marginalized developers) to feel more comfortable in contributing since they know there are guidelines that can protect them from toxic interactions and 2) signal to developers who are already in the community (e.g., including those that may have been inciting toxic interactions) that there will be repercussions for their actions. Unfortunately, less than 10 percent of the top OSS projects actually have one [62]. 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. A risk of this approach is the possibility of project maintainers creating token codes of conduct just to satisfy conditions to receive rewards. This approach should therefore be coupled with evaluation of the code of conduct to ensure that it is both meaningful and actually enforced.

Beyond code of conduct, other potential ways to promote development of friendlier communities are usage of social metrics for community self-evaluation and improvement. A example may be drawn from sites that show employer reviews such as GlassDoor[12] and various job search portals. In OSS context, ability to provide and show contributor reviews as well as other metrics such as distribution of contributor tenure can help developers evaluate potential projects to join, and also provide an OSS project community a means to evaluate what they have or have not done well and how to improve their community.

Challenge: Many communities currently do not have or enforce code of conduct, and aspiring contributors also can't easily evaluate community quality of a given OSS project.

Opportunity: Improvements can be done by promoting creation and usage of codes of conduct across communities, and to provide set of social metrics to help aspiring contributors evaluate quality of community they consider joining.

5.2.2 Mentorship and Highlighting of Role Models

Highlighting of Regional/Women Developers as Role Models. 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 [63], [64]. 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 versus self expression [65]). This experience can take advantage of being in the same shared region by conducting guidance through offline interventions [66]. The duality of fostering both the same community online based on a personal offline experience can further support inclusion.

Another approach that can be used is to highlight role models from underrepresented demographics. For example, our survey results indicate that women developers are interested in mechanisms that highlight the contribution of women. Such mechanisms can be implemented both online

7. https://girlswhocode.com/
8. https://www.womenwhocode.com/
9. https://www.ncwit.org/
10. https://women.acm.org/
11. https://github.blog/2018-11-08-100m-repos/
12. https://www.glassdoor.com/

and offline. Online mechanisms can be in the form of updates to pages such as GitHub Explore [67] to add sections that highlight rising or top developers from underrepresented communities. For offline implementation of this mechanism, developer communities can for example organize and encourage technical presentations and talks by experienced developers from underrepresented demographics.

Challenge: There is lack of mechanism to highlight contribution of developers from underrepresented demographics.

Opportunity: Mechanisms that highlight developers that are popular globally can be augmented to also highlight top or popular developers from more specific demographics.

5.2.3 Diversity Promotion via Automated Software Engineering Tools

Some barriers appear to present opportunities for applying automated software engineering approaches to attract diverse contributors to OSS projects. Existing works [68], [69] highlight the importance of prior social links with existing contributors in developers' decision to join an OSS project, and this can be exploited to promote diversity by augmenting existing approaches with social considerations. We discuss some specific categories of tools in the following paragraphs.

Automated Project Recommenders can be augmented to take into account social considerations. A small number of recent project recommenders [70], [71] factor in developer's social ties, and GitHub itself takes into account which developers a user "follows" when recommending projects in its GitHub Explore [67] page. However, to promote diversity or participation from particular gender/region, these can be further augmented with additional metrics based on recommendations in the survey responses, for example:

- Metrics related to quality of community. For example, typical tenure of contributors (as a proxy of how much contributors enjoy being in the community), reputation of current contributors, and range of current contributors' experience levels (as a proxy of how welcoming the project is to beginners).
- Number of current contributors known to be from similar region as the developer considering to join the project.
- Diversity of current set of active contributors with known gender and/or location.

Automated documentation improvement can be employed more widely to reduce barriers to contribution. This can include application and enhancement of automated document localization techniques to overcome language barriers and support local languages from regions with large numbers of potential contributors. This may be coupled with application of automated techniques to improve readability, completeness and/or quality of artifacts such as README files [72] and release notes [73]. Usage of automated document generation of source code summary [74] and tracking of outdated API names [75] can further reduce time required from potential contributors. This will be valuable especially in regions where OSS projects are more commonly treated as hobby or volunteer work, since reduced time barrier will enable more people to contribute even without monetary rewards.

Automated developer assignment mechanisms can be updated to distribute exciting/challenging tasks more widely to motivate continued participation. This may be in form of modification to existing automated bug assignment techniques such as [76] and [77], that currently are usually used to speed up resolution process [78] instead of to spreading interesting tasks to team members.

Challenge: Current automated software engineering tools tend to focus on technical aspects and similarity between developers (homophily) when making recommendations.

Opportunity: There's opportunity to augment existing tools to enable selection of target social objectives, such as maintenance of contributor interest (by making more even distribution of challenging tasks) or encouraging participation from certain underrepresented communities.

6 Threats to Validity

Construct Validity. Our study has two parts: a large scale data analysis and a survey. During the study design, we made choices that can potentially influence the outcome. Regarding repository selection, the filtering criteria we use still leaves some possibility of including repositories of academic projects that run beyond 6 months, however, we believe that those are also likely to be a more serious endeavor instead of simple programming assignments. Another factor is the accuracy of gender and location resolution. While many factors can cause incorrect gender and location resolution (e.g., incorrect information on GitHub profile, decision to make accounts private), we tried mitigating this threat in two ways. First, we choose a tool that has reportedly reasonable accuracy for multiple regions such as Asia and Eastern Europe [30], [31] and has been used in various studies related to gender representation [32], [33], [34]. Prior to full-scale analysis, we also performed validation by manually checking a subset of the data to increase our confidence in the gender prediction. We also limited our analysis to commit authors, who are more likely to be a code-contributing part of the project team (compared to, for example, issue reporters) and are also more likely to provide information which can be used to resolve their gender and location. Finally, we eliminated projects for whom we could not infer gender and location of at least 75 percent of commit authors. While it is also possible to perform additional validation after the survey by comparing self-reported gender and geography in the response to the information inferred from data analysis, we did not do so as we did not ask prior permission from survey participants for such data usage. This is in compliance with the GDPR and broader research ethical considerations.

We also note that the tool we use (*genderize.io*) is not reflective of a broad gender spectrum. While analysis of nonbinary identities is a research challenge that has received increasing research attention [79], [80], we are currently unaware of methods toreliably assess this in software

systems at a large scale. Future research should investigate this deeper. As none of our survey respondents identified themselves as non-binary, we believe this limitation of genderize.io does not propose a significant threat to the validity of our subsequent analyses.

With respect to our survey, the underrepresentation of women and a broader set of commit authors poses a threat to validity. We attempted to mitigate this by using stratified survey sampling based on gender and location, instead of performing a random sampling of the entire population. For focused survey responses, we asked each participant questions relating to a specific project which we hope provide more concrete response based on the participant's own experience, although there is still some validity risk if the participant has not worked on the project recently.

Internal Validity. Our analysis indicates regional and gender-based differences for open source participants on GitHub. To improve the internal validity of our data analysis, we calculated diversity at different times using two metrics. Our results point in the same direction. Likewise, our survey borrows elements from literature (corroborating with its findings) and builds on it. Using strategic sampling techniques we tried to gather a representative sample to offer a worldwide view.

External Validity. The representatives of our findings is defined by the range of software projects studied. We selected a wide variety of software projects, nevertheless, we might have systematically missed projects which did not meet our prerequisites (e.g., infer gender and location).

Likewise, due to our methodology and scope of respondents at the intersection of both marginalized genders and underrepresented countries in OSS, we miss the opportunity to provide broad insight into the challenges of having an intersectional identity [81]. Further intersectional methodologies and frameworks should be adopted to explore and amplify the voices of developers in the margins.

7 Conclusion and Future Work

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 is low diversity across regions worldwide, and although there is some variation among regional diversity, the difference is not substantial. 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.

Finally, we also believe it will be helpful if researchers from the different parts of the world perform more in-depth study of gender differences in their own regions. We believe that with better understanding of and connections with local developer communities (including developers who are not active on GitHub), local researchers will likely be able to collect more responses. Further, they will also be able to customize their survey to better focus on any region-specific issues they are aware of.

Acknowledgments

We would like to thank our survey respondents for providing the irperspectives for this study. This research was supported by and the Research Lab for Intelligent Software Engineering (RISE) Operational Fund from the School of Computing and Information Systems (SCIS) at Singapore Management University (SMU).

Dataset Availability

In the interest of encouraging others to replicate and build upon our work, we are sharing our data. The data for this study can be found at: doi: https://doi.org/10.5281/zenodo.4637095.

References

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

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

[3] E. Commission, "Increase in gender gap in the digital sector–Study on women in the digital age," 2018. [Online]. Available: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50224

[4] 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 *Proc. 14th ACM/IEEE Int. Symp. Empirical Softw. Eng. Measure.*, 2020, pp. 1–11.

[5] 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," *J. Softw. Eng. Res. Develop.*, vol. 5, no. 1, pp. 1–18, 2017.

[6] A. Bosu and K. Z. Sultana, "Diversity and inclusion in open source software (OSS) projects: Where do we stand?," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Measure.*, 2019, pp. 1–11.

[7] 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," *Manage. Sci.*, vol. 52, no. 7, pp. 984–999, 2006.

[8] 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 *Proc. 18th ACM Conf. Comput. Supported Cooperative Work Social Comput.*, 2015, pp. 1379–1392.

[9] F. Zlotnick, "GitHub open source survey 2017," 2017. [Online]. Available: http://opensourcesurvey.org/2017/

[10] S. Beecham, N. Baddoo, T. Hall, H. Robinson, and H. Sharp, "Motivation in software engineering: A systematic literature review," *Inf. Softw. Technol.*, vol. 50, no. 9–10, pp. 860–878, 2008.

[11] A. C. C. França, T. B. Gouveia, P. C. Santos, C. A. Santana, and F. Q. da Silva, "Motivation in software engineering: A systematic review update," in *Proc. 15th Annu. Conf. Eval. Assessment Softw. Eng.*, 2011, pp. 154–163.

[12] C. França, F. Q. Da Silva, and H. Sharp, "Motivation and satisfaction of software engineers," *IEEE Trans. Softw. Eng.*, vol. 46, no. 2, pp. 118–140, Feb. 2020.

[13] S. O. Alexander Hars , "Working for free? Motivations for participating in open-source projects," *Int. J. Electron. Commerce*, vol. 6, no. 3, pp. 25–39, 2002.

[14] K. R. Lakhani and R. G. Wolf, "Why hackers do what they do: Understanding motivation and effort in free/open source software projects," 2003. [Online]. Available: https://ssrn.com/abstract=443040

[15] G. Von Krogh, S. Haefliger, S. Spaeth, and M. W. Wallin, "Carrots and rainbows: Motivation and social practice in open source software development," *MIS Quart.*, vol. 36, pp. 649–676, 2012.

[16] A. Barcomb, K.-J. Stol, D. Riehle, and B. Fitzgerald, "Why do episodic volunteers stay in floss communities?," in *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng.*, 2019, pp. 948–959.

[17] M. Gerosa *et al.*, "The shifting sands of motivation: Revisiting what drives contributors in open source," 2021, *arXiv:2101.10291*.

[18] C. Mendez *et al.*, "Open source barriers to entry, revisited: A sociotechnical perspective," in *Proc. 40th Int. Conf. Softw. Eng.*, 2018, pp. 1004–1015.

[19] J. Terrell *et al.*, "Gender differences and bias in open source: Pull request acceptance of women versus men," *PeerJ Comput. Sci.*, vol. 3, p. e111, May 2017.

[20] A. Rastogi, N. Nagappan, G. Gousios, and A. van der Hoek, "Relationship between geographical location and evaluation of developer contributions in Github," in *Proc. 12th ACM/IEEE Int. Symp. Empirical Softw. Eng. Measure.*, 2018, pp. 1–8.

[21] D. Ford, R. Milewicz, and A. Serebrenik, "How remote work can foster a more inclusive environment for transgender developers," in *Proc. IEEE/ACM 2nd Int. Workshop Gender Equality Softw. Eng.*, 2019, pp. 9–12.

[22] I. Steinmacher, T. U. Conte, C. Treude, and M. A. Gerosa, "Overcoming open source project entry barriers with a portal for newcomers," in *Proc. 38th Int. Conf. Softw. Eng.*, 2016, pp. 273–284.

[23] B. Vasilescu, V. Filkov, and A. Serebrenik, "Perceptions of diversity on GitHub: A user survey," in *Proc. IEEE/ACM 8th Int. Workshop Cooperative Hum. Aspects Softw. Eng.*, 2015, pp. 50–56.

[24] G. Robles, L. A. Reina, J. M. Gonz ález-Barahona, and S. D. Domínguez, "Women in free/libre/open source software: The situation in the 2010s," in *Proc. IFIP Int. Conf. Open Source Syst.*, 2016, pp. 163–173.

[25] B. Vasilescu *et al.*, "Gender and tenure diversity in GitHub teams," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 3789–3798.

[26] A. Lee and J. C. Carver, "Floss participants' perceptions about gender and inclusiveness: A survey," in *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng.*, 2019, pp. 677–687.

[27] N. Imtiaz, J. Middleton, J. Chakraborty, N. Robson, G. Bai, and E. Murphy-Hill, "Investigating the effects of gender bias on GitHub," in *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng.*, 2019, pp. 700–711.

[28] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. London, U.K.: SAGE, 2018.

[29] S. Zhou, S. Stanciulescu, O. Leßenich, Y. Xiong, A. Wasowski, and C. Kästner, "Identifying features in forks," in *Proc. IEEE/ACM 40th Int. Conf. Softw. Eng.*, 2018, pp. 105–116.

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

[31] F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, and M. Strohmaier, "Inferring gender from names on the web: A comparative evaluation of gender detection methods," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 53–54.

[32] L. Holman, D. Stuart-Fox , and C. E. Hauser, "The gender gap in science: How long until women are equally represented?," *PLoS Biol.*, vol. 16, no. 4, p. e2004956, 2018.

[33] E. G. Thomas, B. Jayabalasingham, T. Collins, J. Geertzen, C. Bui, and F. Dominici, "Gender disparities in invited commentary authorship in 2459 medical journals," *JAMA Netw. Open*, vol. 2, no. 10, pp. e1913682–e1913682, 2019.

[34] P. H. Russell, R. L. Johnson, S. Ananthan, B. Harnke, and N. E. Carlson, "A large-scale analysis of bioinformatics code on GitHub," *PLoS One*, vol. 13, no. 10, p. e0205898, 2018.

[35] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, and D. Damian, "The promises and perils of mining GitHub," in *Proc. 11th Working Conf. Mining Softw. Repositories*, 2014, pp. 92–101.

[36] P. M. Blau, *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York, NW, USA: Free Press, 1977, vol. 7.

[37] B. Vasilescu, A. Serebrenik, and V. Filkov, "A data set for social diversity studies of GitHub teams," in *Proc. IEEE/ACM 12th Working Conf. Mining Softw. Repositories*, 2015, pp. 514–517.

[38] 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 *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng.: Softw. Eng. Soc.*, 2019, pp. 11–20.

[39] H. Cramer, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton Univ. Press, 1946.

[40] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, pp. 50–60, 1947.

[41] H. Abdi, "Bonferroni and šidák corrections for multiple comparisons," *Encyclopedia Measure. Statist.*, vol. 3, pp. 103–107, 2007.

[42] N. Cliff, *Ordinal Methods for Behav. Data Anal*. Hove, U.K.: Psychology Press, 2014.

[43] B. Trinkenreich, M. Guizani, I. Wiese, A. Sarma, and I. Steinmacher, "Hidden figures: Roles and pathways of successful OSS contributors," in *Proc. ACM Hum.-Comput. Interact.*, 2020, pp. 1–22.

[44] R. B. D'agostino, A. Belanger, and R. B. D'Agostino Jr, "A suggestion for using powerful and informative tests of normality," *Amer. Statist.*, vol. 44, no. 4, pp. 316–321, 1990.

[45] W. G. Hopkins, *A new view of statistics*. Will G. Hopkins, 1997.

[46] P. Virtanen *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, pp. 261–272, 2020.

[47] F. Camilo, A. Meneely, and M. Nagappan, "Do bugs foreshadow vulnerabilities? A study of the chromium project," in *Proc. IEEE/ACM 12th Working Conf. Mining Softw. Repositories*, 2015, pp. 269–279.

[48] GitHub, "GitHub milestones: A timeline of significant moments in GitHub's history," 2020, Accessed: May 28, 2020. [Online]. Availalbe: https://github.com/about/milestones

[49] 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 *Proc. IEEE/ACM 39th Int. Conf. Softw. Eng.*, 2017, pp. 187–197.

[50] D. Ford, J. Smith, P. J. Guo, and C. Parnin, "Paradise unplugged: Identifying barriers for female participation on stack overflow," in *Proc. 24th ACM SIGSOFT Int. Symp. Found. Softw. Eng.*, 2016, pp. 846–857.

[51] D. Ford, A. Harkins, and C. Parnin, "Someone like me: How does peer parity influence participation of women on stack overflow?," in *Proc. IEEE Symp. Vis. Lang. Hum.-Centric Comput.*, 2017, pp. 239–243.

[52] J. Saldaña, *The Coding Manual for Qualitative Researchers*. London, U.K.: SAGE, 2009.

[53] J. L. Campbell, C. Quincy, J. Osserman, and O. K. Pedersen, "Coding in-depth semistructured interviews," *Sociol. Methods Res.*, vol. 42, no. 3, pp. 294–320, Aug. 2013.

[54] N. A. Ibe, R. Howsmon, L. Penney, N. Granor, L. A. DeLyser, and K. Wang, "Reflections of a diversity, equity, and inclusion working group based on data from a national cs education program," in *Proc. 49th ACM Tech. Symp. Comput. Sci. Educ.*, 2018, pp. 711–716.

[55] M. E. Vachovsky, G. Wu, S. Chaturapruek, O. Russakovsky, R. Sommer, and L. Fei-Fei , "Toward more gender diversity in CS through an artificial intelligence summer program for high school girls," in *Proc. 47th ACM Tech. Symp. Comput. Sci. Educ.*, 2016, pp. 303–308.

[56] V. Borsotti, "SIFSOFT distinguished paper–Barriers to gender diversity in software development education: Actionable insights from a Danish case study," in *Proc. IEEE/ACM 40th Int. Conf. Softw. Eng.: Softw. Eng. Educ. Training.*, 2018, pp. 146–152.

[57] A. Kulkarni, I. Yoon, P. S. Pennings, K. Okada, and C. Domingo, "Promoting diversity in computing," in *Proc. 23rd Annu. ACM Conf. Innov. Technol. Comput. Sci. Educ.*, 2018, pp. 236–241.

[58] P. Rheingans, E. D'Eramo, C. Diaz-Espinoza , and D. Ireland, "A model for increasing gender diversity in technology," in *Proc. 49th ACM Tech. Symp. Comput. Sci. Educ.*, 2018, pp. 459–464.

[59] T. Kelly and R. Firestone, "How tech hubs are helping to drive economic growth in africa," World Develop. Rep., World Bank, Washington, DC, USA, Tech. Rep. TR-102957, 2016.

[60] V. Singh, "Women participation in open source software communities," in *Proc. 13th Eur. Conf. Softw. Architect.*, 2019, pp. 94–99.

[61] C. A. Ehmke, "Contributor covenant," 2014, Accessed: Mar. 05, 2020. [Online]. Available: https://www.contributor-covenant.org/

[62] 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, Switzerland: Springer, 2019, pp. 68–79.
[63] Z. Brodsky, "7 Chinese open source projects you should know about," 2018, Accessed: Mar. 05, 2020. [Online]. Available: https://resources.whitesourcesoftware.com/blog-whitesource/7-chinese-op en-source-projects-you-should-know-about
[64] A. Abati, "Made in Nigeria," 2017, Accessed: Mar. 05, 2020. [Online]. Available: https://www.madeinnigeria.dev/
[65] N. Oliveira, N. Andrade, and K. Reinecke, "Participation differences in Q&A sites across countries: Opportunities for cultural adaptation," in *Proc. 9th Nordic Conf. Hum.-Comput. Interact.*, 2016, pp. 1–10.
[66] Debian, "Debian women mentoring program," 2019, Accessed: Mar. 05, 2020. [Online]. Available: https://www.debian.org/women/mentoring
[67] GitHub, "Explore GitHub," 2021, Accessed: Mar. 19, 2021. [Online]. Available: https://github.com/explore
[68] C. Casalnuovo, B. Vasilescu, P. Devanbu, and V. Filkov, "Developer onboarding in GitHub: The role of prior social links and language experience," in *Proc. 10th Joint Meeting Found. Softw. Eng.*, 2015, pp. 817–828.
[69] J. Hahn, J. Y. Moon, and C. Zhang, "Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties," *Inf. Syst. Res.*, vol. 19, no. 3, pp. 369–391, 2008.
[70] C. Liu, D. Yang, X. Zhang, B. Ray, and M. M. Rahman, "Recommending GitHub projects for developer onboarding," *IEEE Access*, vol. 6, pp. 52 082–52 094, Sep. 2018.
[71] T. Matek and S. T. Zebec, "GitHub open source project recommendation system," 2016, *arXiv:1602.02594*.
[72] G. A. A. Prana, C. Treude, F. Thung, T. Atapattu, and D. Lo, "Categorizing the content of GitHub readme files," *Empirical Softw. Eng.*, vol. 24, no. 3, pp. 1296–1327, 2019.
[73] L. Moreno, G. Bavota, M. Di Penta , R. Oliveto, A. Marcus, and G. Canfora, "Automatic generation of release notes," in *Proc. 22nd ACM SIGSOFT Int. Symp. Found. Softw. Eng.*, 2014, pp. 484–495.
[74] P. W. McBurney and C. McMillan , "Automatic documentation generation via source code summarization of method context," in *Proc. 22nd Int. Conf. Program Comprehension*, 2014, pp. 279–290.
[75] S. Lee, R. Wu, S.-C. Cheung, and S. Kang, "Automatic detection and update suggestion for outdated API names in documentation," *IEEE Trans. Softw. Eng.*, vol. 47, no. 4, pp. 653–675, Apr. 2021.
[76] L. Jonsson, M. Borg, D. Broman, K. Sandahl, S. Eldh, and P. Runeson, "Automated bug assignment: Ensemble-based machine learning in large scale industrial contexts," *Empirical Softw. Eng.*, vol. 21, no. 4, pp. 1533–1578, 2016.
[77] T. Zhang, J. Chen, H. Jiang, X. Luo, and X. Xia, "Bug report enrichment with application of automated fixer recommendation," in *Proc. IEEE/ACM 25th Int. Conf. Program Comprehension*, 2017, pp. 230–240.
[78] W. Zou, D. Lo, Z. Chen, X. Xia, Y. Feng, and B. Xu, "How practitioners perceive automated bug report management techniques," *IEEE Trans. Softw. Eng.*, vol. 46, no. 8, pp. 836–862, Aug. 2020.
[79] F. Hamidi, M. K. Scheuerman, and S. M. Branham, "Gender recognition or gender reductionism? The social implications of embedded gender recognition systems," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–13.
[80] O. Keyes, "The misgendering machines: Trans/HCI implications of automatic gender recognition," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–22, 2018.
[81] Y. A. Rankin and J. O. Thomas, "Straighten up and fly right: Rethinking intersectionality in HCI research," *Interactions*, vol. 26, no. 6, p. 64–68, Oct. 2019.

Gede Artha Azriadi Prana is currently working toward the PhD degree with the School of Computing and Information Systems, Singapore Management University, under the supervision of Professor David Lo. Prior to his current study, he was in the field of software engineering for about a decade. His research interest focuses on software engineering analytics.

Denae Ford received the PhD degree in computer science from North Carolina State University. She is currently a senior researcher with Microsoft Research, where she works with the Software Analysis and Intelligence team and is an affiliate assistant professor with the Human-Centered Design and Engineering Department, University of Washington. 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 interests include investigating and devising intervention to support inclusive socio-technical interactions in online programming communities.

Ayushi Rastogi studies human and social aspects of software engineering for improving developer productivity and promoting diversity and inclusion. She is currently an assistant professor with the Faculty of Science and Engineering, University of Groningen, The Netherlands. Her research interests include software analytics, empirical software engineering, and mining software repositories.

David Lo received the PhD degree in computer science from the National University of Singapore. He is currently a professor with the School of Computing and Information Systems, Singapore Management University. His research interests include software analytics, software maintenance, software testing, social network mining, and cybersecurity.

Rahul Purandare received the PhD degree in computer science from the University of Nebraska - Lincoln. He is currently an associate professor with the Department of Computer Science and Engineering, Indraprastha Institute of Information Technology Delhi. His research interests include program analysis, software testing, automatic program repair, code search, and code comprehension.

Nachiappan Nagappan (Fellow, IEEE) was a partner researcher with Microsoft Research when this work was performed. He also holds an adjunct faculty appointment with the Indraprastha Institute of Information Technology Delhi. His research interests include 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. He is a Fellow of the ACM.