

An Empirical Comparison of Ethnic and Gender Diversity of DevOps and non-DevOps Contributions to Open-Source Projects

Nimmi Rashinika Weeraddana · Xiaoyan Xu · Mahmoud Alfadel · Shane McIntosh · Meiyappan Nagappan

Author pre-print copy. The final publication is available at Springer via: <https://doi.org/10.1007/s10664-023-10394-9>

Abstract Diversity has been recognized as a high-value team characteristic. Both open-source and proprietary software organizations have been investing heavily in creating more diverse teams. Prior work has raised diversity concerns about open-source communities; however, to the best of our knowledge, it is not yet clear if those diversity concerns permeate across all of the subteams of the project. Studying diversity in subteams would provide more detailed empirical evidence about the role of diversity in software development teams.

Therefore, we perform an empirical study on 110,336 developers who contributed to artifacts of 450 large and thriving open-source projects. We opt to study diversity of the DevOps team because it plays a central role in a project. In particular, we analyze the perceptible ethnic and gender diversity among DevOps contributors to open-source, and we ground our analysis in a comparison to non-DevOps contributors. Overall, our results show that, with respect to perceptible ethnic diversity, contributors with perceptibly White names in a project are the majority of DevOps contributors (median = 87.70%) and non-DevOps contributors (median = 85.50%). With respect to gender diversity, contributors who are perceptible as men in a project are the majority of DevOps contributors (median = 93.75%) and non-DevOps contributors (median = 92.82%). We statistically measure the perceptible ethnic and gender diversity of both DevOps and non-DevOps contributors using diversity metrics, and we find that the diversity of DevOps contributors is significantly less than that of non-DevOps contributors. When analyzing the distribution of diversity change as projects evolve, we find that contributors perceptible as non-Whites (such as Hispanic and Black) are greatly underrepresented. Although the percentage of contributors perceptible as White is decreasing over time, the percentage of contributors perceptible as non-White is still low, i.e., it varies between 0%–16.02% for DevOps and

Nimmi Rashinika Weeraddana, Xiaoyan Xu, Mahmoud Alfadel, Shane McIntosh, Meiyappan Nagappan
University of Waterloo, Canada
E-mail: {nrweeraddana,xiaoyan.xu,malfadel,shane.mcintosh,mei.nagappan}@uwaterloo.ca

0%–18.77% for non-DevOps. We observe similar results for gender diversity, where contributors perceptible as men dominate over contributors perceptible as women.

Our study provides empirical evidence contributing towards a better understanding of diversity aspects from a different perspective (DevOps vs. non-DevOps contributors). Our findings call for higher awareness, not only of the overall diversity but also of the diversity in specific subteams of the project.

Keywords DevOps, Diversity, Ethnicity, Gender, Diversity metrics

1 Introduction

Diversity in any profession benefits everyone involved and can lead to a large positive impact on different communities [24, 55]. More specifically, in the software engineering community, greater diversity has been associated with more resilient software solutions, problem-solving from broader perspectives, and tends to improve the effectiveness of teamwork [72]. At present, large organizations, such as Google,¹ Microsoft,² and Facebook³ are investing heavily in fostering diverse teams.

Previous work has explored some aspects of diversity in open-source communities [6, 10, 46, 70, 72]. For example, Nadri et al. [46] empirically analyzed the impact of race and ethnicity on the likelihood of acceptance of the proposed contributions in open-source projects. They showed that contributors with perceptibly White names are responsible for the vast majority of the contributions, and their contributions are more likely to be accepted than the contributions of contributors with perceptibly non-White names. Vasilescu et al. [72] observed that contributors who are perceptible as women account for a smaller proportion of GitHub project teams than men. In addition, they found that teams with higher rates of gender diversity tend to be more productive.

While diversity in software development teams has been extensively studied [10, 70, 72], much of this research has focused on general investigations of diversity without specific attention to particular types of subteams within teams, e.g., quality assurance teams and DevOps teams. Studying diversity in subteams would provide more precise and targeted empirical evidence about the role of diversity in software development teams, as it would allow us to examine whether the diversity concerns that have been observed overall permeate all of the subteams of a project. For instance, developers in the DevOps team of a project require a specialized skill set [77]. In particular, DevOps contributors typically require expertise in collaboration, automation, measurement, monitoring [34], and Agile methodologies [3, 4], as well as experience with cloud-based tools and infrastructure [3, 12]. Such specialized skills and tooling expertise may not be as common in the wider open-source community. This is evident in the Stack Overflow survey (2022),⁴ as it shows that only 10.06% of respondents identified themselves as DevOps developers. The survey results also reveal a gap in

¹ <https://diversity.google/>

² <https://www.microsoft.com/en-us/diversity>

³ <https://diversity.fb.com/read-report/>

⁴ <https://survey.stackoverflow.co/2022/>

the gender diversity of DevOps developers.⁵ Similar diversity concerns have been raised in popular DevOps reports as well. For example, DORA's *State of DevOps report (2021)* [64] highlighted that the percentage of DevOps developers who found themselves underrepresented has increased from 13.7% in 2019 to 17% in 2021.

Therefore, we set out to study the diversity among DevOps contributors to open-source projects. Note that, to the best of our knowledge, this is the first study on the diversity of subteams within project teams, and we opt to focus on DevOps teams because DevOps plays a central role in a project [31]. In particular, the DevOps contributors in a project are the ones who are responsible for supporting other teams during software development, continuous integration [75], continuous product delivery, and managing and allocating resources, such as computing and storage, to other teams [18, 41]. Diversity in DevOps can help to improve the overall performance of a project. In fact, a report, "Why diversity matters in DevOps,"⁶ discusses the potential of diversity in DevOps teams to lead to more streamlined product development. Moreover, there is a growing focus on diversity of DevOps contributors in the industry as well. For example, the global movement "Women in DevOps"⁷ aims to close the gender gap in DevOps contributors, believing that a balanced and diverse workforce drives innovation. We ground our analysis of diversity metrics of developers in DevOps teams in comparison to other developers. More specifically, we perform an empirical study involving data from 450 active and mature projects that are hosted on GitHub.⁸ A preliminary analysis of this data indicates that the majority (median = 86.70%) of the developers in a project concentrate their contributions on non-DevOps artifacts only. Yet, a non-negligible percentage (median = 13.30%) of developers contribute to DevOps artifacts. Therefore, we formulate and address the following research questions:

(RQ1) Does the perceptible ethnic and gender diversity of DevOps contributors differ from ethnic and gender diversity of non-DevOps contributors?

Motivation: While there is a lack of research specifically focused on the impact of diversity in DevOps teams, there is evidence regarding the positive impact of diversity in software development teams [15, 33, 72]. Since DevOps plays a major role in the software development process [16], a diverse DevOps team would ensure the DevOps process is more robust, inclusive, and successful. Therefore, this RQ aims to provide the community with more awareness of the presence or lack of ethnic and gender diversity of DevOps contributors to open-source projects.

Results: With respect to ethnic diversity, contributors with perceptibly White names are the majority among both DevOps contributors (median = 87.70%) and non-DevOps contributors (median = 85.50%). With respect to gender diversity, contributors with names perceived as men are the majority among

⁵ Stack Overflow survey (2022) shows that the percentage of women among the DevOps developers (2.10%) is less than that of other women developers (5.13%).

⁶ <https://www.harrisonclarke.com/devops-sre-recruiting-blog/diversity-and-inclusion-in-devops>

⁷ <https://www.womenindevops.com/>

⁸ <https://github.com>

both DevOps contributors (median = 93.75%) and non-DevOps contributors (median = 92.82%). We statistically measure the ethnic and gender diversity of both DevOps and non-DevOps contributors using diversity metrics, and we find that diversity of DevOps contributors is significantly less than that of non-DevOps contributors (Wilcoxon signed rank test, $p < \alpha = 0.0023$,⁹ one-tailed, paired).

(RQ2) How does the distribution of perceptible ethnic and gender diversity change as projects age?

Motivation. While a concerning lack of ethnic and gender diversity in open-source communities has been reported for decades now [13, 21, 76], it is not yet clear where the current trend is headed. By analyzing the diversity metrics over time, we can better understand whether the trend of diversity is improving or further degrading. Therefore, we examine the evolution of ethnic and gender diversity of DevOps and non-DevOps contributors in the projects.

Results: As projects evolve, contributors perceptible as non-White remain greatly underrepresented. The overall percentage of DevOps contributors perceptible as White is decreasing over time; however, the percentage of DevOps contributors perceptible as non-White is still low, i.e., it varies between 0%-16.02%. For non-DevOps contributors, the percentage varies between 0%-18.77%. With respect to gender diversity, both DevOps and non-DevOps contributors who are perceptible as men dominate over contributors who are perceptible as women. The percentage of contributors perceptible as women varies between 0%-12.5% and between 0%-9.48% for DevOps and non-DevOps, respectively.

The above answers to our research questions revealed that minorities with respect to perceived ethnicity and gender are more underrepresented among DevOps contributors compared to non-DevOps contributors. However, Ross et al. [57] discovered that developers at the intersection of being Black and being a woman (i.e., Black women) often have different experiences than Black men and non-Black women in the US. In our study, independently analyzing ethnic and gender diversity may not help to understand how diverse the contributors are when considering the intersection of minority groups of perceptible ethnicity and gender (e.g., contributors with perceptibly non-White and women's names). To that end, we perform an intersectional analysis of ethnic and gender diversity to obtain a more realistic picture with regard to ethnicity and gender [29]. Our results show that perceptible Whites are the majority of DevOps and non-DevOps contributors who are perceptible as women. More importantly, we find that DevOps contributors' lack of perceived diversity is amplified when considering intersectionality.

We believe that our study provides empirical evidence that contributes towards a better understanding of the perceptible diversity among contributors in open source. We complement prior studies that raise awareness of the lack of perceptible diversity among contributors in open-source projects. While solutions and strategies have

⁹ The Bonferroni-corrected significance level in our study is 0.0023.

been proposed to increase diversity in open-source projects ([13, 21, 76]), our results underscore the importance of encouraging open-source communities to foster a more diverse and inclusive environment, not only considering the overall project team but also the different subteams within the project.

Data availability statement. The dataset we curated and analyzed during the current study is available online.¹⁰

Paper organization. The remainder of the paper is organized as follows. Section 2 provides an overview of the design of our study. Section 3 presents the results for our research questions. Section 4 discusses our results further. We present the threats to the validity of the study in Section 5. Section 6 discusses the related work to our study. Finally, Section 7 draws conclusions and the broader implications of our study.

2 Study Design

In this section, we describe our process for collecting and curating the dataset we use to address our research questions. Fig. 1 provides an overview of our study design, which is composed of Project Selection (PS), Data Curation (DC), and Preliminary Analysis (PA) steps. Next, we explain each step in detail.

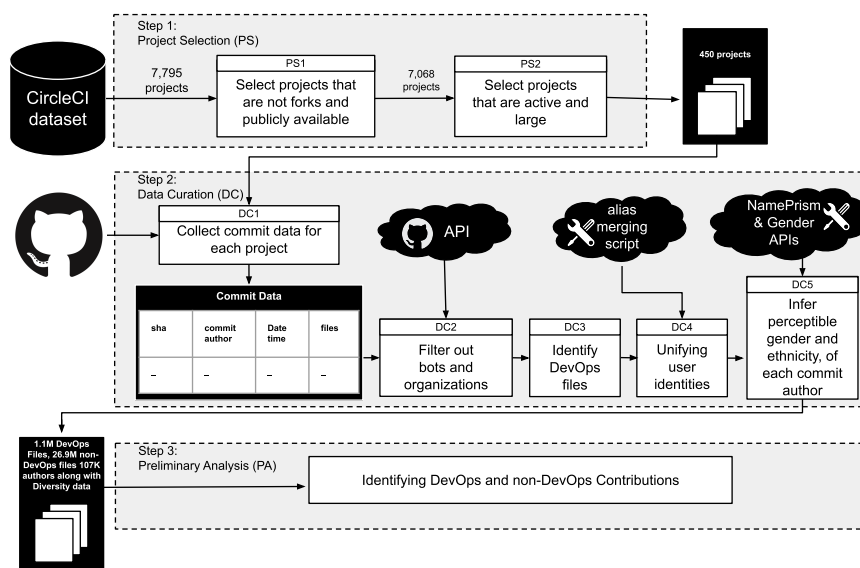


Fig. 1: An overview of our study design showing Project Selection (PS), Data Curation (DC), and Preliminary Analysis (PA) steps.

¹⁰ <https://doi.org/10.5281/zenodo.8277702>

(PS) Project Selection

Our study aims to analyze the diversity of contributors to DevOps and non-DevOps artifacts in open-source projects. Thus, we need to collect a dataset of open-source projects that adopt tools and technologies for DevOps activities. We begin with the public dataset of Gallaba et al. [25]. This dataset contains data from 23,330,690 software builds that span 7,795 GitHub projects that have been using CircleCI,¹¹ a leading cloud-based Continuous Integration (CI) platform that has served over one million unique contributors during its nine years of operation.¹² This dataset ensures that the projects have potentially been using DevOps tools and technologies to help them automate software development processes, such as building, testing, and deploying the software.

Since GitHub hosts projects that are not representative of projects we aim to investigate (e.g., toy or immature projects) [44], we follow the methodology recommended by previous work [38] to further curate our dataset by applying the following inclusion criteria:

(PS1) Select non-forked projects. We remove forks¹³ because they largely contain duplicated project history, which would bias our analysis. To do so, we use the GitHub API¹⁴ to determine whether a project is a fork or not. If the project is a fork, the GitHub API returns the `fork` status `True`, and we filter out all such projects. This step reduces our dataset to 7,068 projects.

(PS2) Select active and large projects. Active and large projects are likely to showcase a long-running and collaborative software development process to examine diversity. To detect active and large projects, we consider different thresholds of (1) the number of builds, (2) the number of commits, and (3) the number of contributors.

Number of builds. Fig. 2 plots candidate threshold values of the number of builds against the number of surviving projects. We select a threshold of 500 builds because it is closer to a “knee” in the curve.¹⁵ Selecting this threshold further reduces the number of projects in the dataset to 2,124.

Number of commits. Fig. 3 plots candidate threshold values of the number of commits against the number of surviving projects. We select a threshold of 1,500 commits because it is also closer to a “knee” in the curve. Doing so reduces the number of projects in the dataset to 850 projects.

Number of contributors. Fig. 4 plots candidate threshold values for the number of contributors against the number of surviving projects. We select a threshold of 50 contributors because we wanted to study projects with a substantial number of contributors. Doing so reduces the number of projects in our dataset to 450.

¹¹ <https://circleci.com/>

¹² <https://circleci.com/milestones/>

¹³ A fork is a copy of a project that a GitHub user manages. Forks let GitHub users make changes to a project without affecting the original project.

¹⁴ <https://docs.github.com/en/rest/repos/repos#get-a-repository>

¹⁵ Note that we do not use the Kneedle algorithm[61] (a knee detecting algorithm) to detect knees because such algorithms use a conservative approach to detect knees by identifying a clear inflection point, resulting in an overly conservative number of surviving projects.

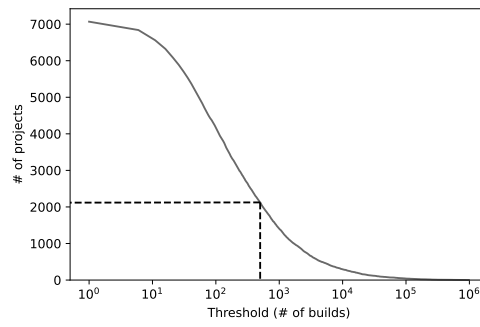


Fig. 2: Threshold plot for the number of builds.

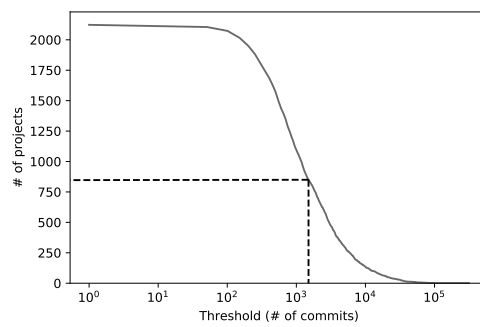


Fig. 3: Threshold plot for the number of commits.

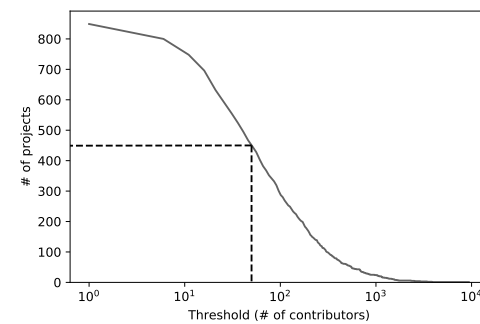


Fig. 4: Threshold plot for the number of contributors.

This dataset of 450 projects comprises large projects from popular organizations, e.g., Facebook,¹⁶ Google,¹⁷ and Angular.¹⁸ Overall, the projects in our dataset run a

¹⁶ <https://github.com/facebook>

considerable number of builds (a median of 8,711 builds per project) and have a rich development history (a median of 4,935 commits per project and 145 contributors per project).

(DC) Data Curation

In our study, we integrate data from various sources to acquire all the data we need for our analysis. Fig. 1 (Step 2 DC) provides an overview of our data curation process. In the following, we describe each data curation step in detail.

(DC1) Collecting commit information.

To analyze the diversity of contributors, we need to collect historical data on project development activity. Therefore, we first clone local copies of the 450 repositories in our dataset. Then, for each repository, we mine the commit records that appear on its `master/main` branch (doing so ensures that the most impactful changes to the source code are considered and mitigates the inconsistencies that may arise due to the deletion of temporary branches). For each commit, we extract (meta)data, such as the unique commit ID (i.e., SHA), timestamp, name of the contributor who authored the commit, and files modified (or created) in the commit.

(DC2) Filtering out commits from non-human contributors. Our goal is to examine ethnic and gender diversity among contributors in open-source projects, and hence, any contribution made by non-humans (e.g., bots) should not be considered in our data analysis. Therefore, we filter out commits made by non-human contributors. To do so, we use the GitHub API¹⁹ to determine the commit author type, i.e., if the commit is made by a human contributor, the GitHub API returns the type `USER`. We filter out all commits made by authors of a different type. For example, contributions associated with the type `BOT` indicate that the commit is made by a bot (e.g., Dependabot).

(DC3) Identifying DevOps artifacts. Since we aim to analyze the diversity of contributors to DevOps artifacts, we need to identify potential DevOps artifacts in the studied projects. To do so, we first search for DevOps tools that have the potential to be used in the studied projects. Then, we identify the adoption of such tools in the projects.

Selecting DevOps tools. We opt to select the tools that are used for CI, Deployment, Containers, Builds, and Configuration because such DevOps practices are commonly used in projects that perform DevOps [58]. The table of DevOps tools [1] offered by Digital.ai²⁰ contains a list of popular DevOps tools used for these prac-

¹⁷ <https://github.com/google/mtail>

¹⁸ <https://github.com/angular/angular>

¹⁹ <https://docs.github.com/en/rest/users/users#about-the-users-api>

²⁰ <https://digital.ai/>

tices. For instance, for CI, the listed tools are CircleCI,²¹ TravisCI,²² Jenkins,²³ AWS CodeBuild,²⁴ and Codeship.²⁵

Note that there are several other websites that list DevOps tools (e.g., Atlassian,²⁶ Guru,²⁷ Geekflare,²⁸ etc.). We choose the Digital.ai periodic table for two reasons. First, this periodic table offers a comprehensive and organized representation of the DevOps tools [1]. More specifically, it defines 17 categories of tools across five different licensing models from open source through enterprise, covering a wide range of functionalities that cater to diverse IT processes in the DevOps domain. Also, it reflects the votes of over 18,000 DevOps practitioners for over 400 tools. Second, Digital.ai periodic table is widely acknowledged and cited as a reference in academic papers and industry articles in the DevOps field.²⁹

Identifying DevOps files. We inspect the documentation of each tool to identify the filenames that are used to configure those tools. For example, to configure CircleCI, a `.circleci/config.yml` file is required; any project that contains a commit on the `.circleci/config.yml` file is using (or used) CircleCI service.

Some DevOps technologies do not have a filename convention. For example, Ansible³⁰ and Kubernetes³¹ are configured using `.yaml` files, but the team may choose to name the file however they see fit. Since `.yaml` is a popular extension used for all sorts of files, we cannot rely solely on the extension to identify DevOps files. Therefore, the first two authors inspect a sample of 400 `.yaml` files from our dataset that are not explicitly classified as DevOps files based on the filename. The inspectors prepare a list of keywords found in the filename or file path that make a file a *DevOps file* or a *non-DevOps file*. For example, `devops`, `deployment`, `kubernetes`, and `chef` are some keywords in a file path, which indicate that `.yaml` file is a *DevOps file*; in contrast, if the file path contains `changelog`, `docs/`, `package`, or `pullapprove`, then that `.yaml` file is a *non-DevOps file*. During this coding, the inspectors also prepare a list of keywords by inspecting the content of a `.yaml` file. For example, any `.yaml` file related to Kubernetes must contain `kind` and `apiVersion` key settings.

After listing the potential keywords, we identify the `.yaml` files that are likely to be DevOps files by matching keywords to the file paths. For the `.yaml` files that are not easily classifiable, we parse the content in search of key settings that we identified as DevOps-related above. After applying these automatic classification steps, 3.4% of the `.yaml` files remain ambiguous. To ensure the integrity of our categories, we remove the ambiguous files from our analysis.

²¹ <https://circleci.com>

²² <https://www.travis-ci.com>

²³ <https://www.jenkins.io>

²⁴ <https://aws.amazon.com/codebuild>

²⁵ <https://www.cloudbees.com/products/codeship>

²⁶ <https://www.atlassian.com/devops/devops-tools>

²⁷ <https://www.guru99.com/devops-tools.html>

²⁸ <https://geekflare.com/devops-tools/>

²⁹ <https://www.uktech.news/technology-news/digital-ai-releases-new-version-of-industry-standard-periodic-table-of-devops-tools>

³⁰ <https://www.ansible.com>

³¹ <https://kubernetes.io>

After that, we evaluate our classifier. The second author manually classifies a sample of 400 .yml files from our dataset. Then, we calculate the Cohen's Kappa coefficient [11] of agreement between the author and our classifier. This coefficient is commonly used to evaluate inter-rater agreement for categorical items between two raters. The value of Cohen's Kappa coefficient ranges from -1 to +1, with values greater than zero indicating an agreement better than chance. We obtain a Kappa coefficient of 0.82, indicating near-perfect strength of agreement between the author and the classifier. The false positive rate is 0.005, meaning that 0.5% of negative instances are falsely identified as positive, while the false negative rate is 0.227, indicating that 22.7% of positive instances are incorrectly labelled as negative.

Cleaning files dataset. Through a preliminary inspection of the files in our dataset, we observe that a non-negligible proportion is not relevant to the purpose of our analysis, e.g., `node_modules/node-inspector/node_modules/v8-debug/node_modules/node-pre-gyp/node_modules/mkdirp/package.json` and `vendor/k8s.io/kubernetes/cmd/mungedocs/links.go` are automatically generated files that are not maintained by hand. The first two authors inspect a sample of 200 files from our dataset and prepare a list of keywords in a file path that makes a file an auto-generated dependency file or not. For example, the files within the folders named `vendor` and `node_modules` contains third-party dependencies. After listing the potential keywords, we identify the files that are likely to be dependency files by matching keywords to the file paths. The percentage of such files amounts to 15.7% of all the files in our dataset, and we remove them from our analysis to ensure the validity of the results.

To evaluate the aforementioned classifier, the second author manually classifies a sample of 400 files from our dataset. Then, we calculate the Cohen's Kappa coefficient of agreement between the author and our classifier. We obtain a Kappa coefficient of 0.81, showing near-perfect agreement between the author and the classifier. The false positive rate is 0.056, meaning that 5.6% of negative instances are falsely identified as positive, while the false negative rate is 0.047, indicating that 4.7% of positive instances are incorrectly labelled as negative.

Lastly, for each project, we include in our dataset the commits made after the first DevOps-related commit in the project and until the end of the year 2021. Analyzing the commits after the first DevOps-related commit ensures that the timeframes for non-DevOps and DevOps artifacts align.

(DC4) Unifying identities. On GitHub, commits may not be attributed properly due to variations in a committer's name and email address.³² For example, consider the email address `sandyw@gmail.com` that could be associated with two names based on the local configurations of the contributor: `Sandy W` and `Sandy White`. Besides, the same contributor may use different email addresses. For example, `Sandy White` may use their personal email `sandyw@gmail.com` as well as the noreply email address (`sandy@users.noreply.github.com`) as their commit-email address.³³ In order

³² <https://git-scm.com/book/en/v2/Git-Basics-Git-Aliases>

³³ <https://docs.github.com/en/account-and-profile/setting-up-and-managing-your-personal-account-on-github/managing-email-preferences/setting-your-commit-email-address>

to unify such cases with different identities of the same GitHub contributor, we use the `GitHub-alias-merging` script by Vasilescu et al. [73]. This script uses heuristics to link different aliases and email addresses belonging to the same GitHub contributor. By using this script, we identify 110,336 unique contributors from 138,012 `<name, email>` pairs in our original dataset. For the rest of the study, we use these unique contributor identities.

(DC5) Inferring perceptible demographics of contributors. Since we aim to investigate gender and ethnic diversity, we need to classify contributors accordingly. Since gender and ethnic identity are difficult to ascertain at scale, similar to prior work [45, 46, 52, 72], we rely on perceived identity characteristics that can be inferred based on publicly visible profile data.

Inferring perceptible genders. It is not our aim to establish new means of inferring gender. There already exist several open-source tools that are capable of inferring gender based on the name of the contributor, such as `Gender-guesser`,³⁴ `GenderComputer`,³⁵ and `Wiki-Gendersort`.³⁶ Unfortunately, these tools are not without limitations. For example, Santamaría et al. [60] evaluated the `Gender-guesser` tool and found that the tool produced a rate of 20.12% unrecognized names [60]. In such cases, the tool predicts the gender as `None`; for example, the gender predicted from the names, such as `rd`, `ga`, and `xulin` was `None`. To tackle the problem of the high rate of unrecognized names, inspired by prior work [62], we combine the outcomes from all three gender-inferring tools and evaluate the rate of unrecognized names. Among the different combinations we study, the most effective combination to infer gender for the test dataset (provided by Santamaría et al. [60]) is first to infer the gender using `Wiki-Gendersort`, and if the gender is unrecognized, then infer the gender using `Gender-guesser` and `GenderComputer`. Doing so reduces the rate of unrecognized names (20.12%) to 5.15%. Thus, we used the aforementioned combination of tools to infer the perceptible genders of contributors in our dataset. We discard the names that none of the tools are able to infer the perceptible gender from our analysis.

Inferring perceptible ethnicities. Inspired by prior work on diversity [17, 30, 45, 46], we use the `NAME-PRISM` [79] tool to infer the perceptible ethnicity of contributors. `NAME-PRISM` [79] is a name-based perceptible ethnicity classification tool, which uses name-embeddings to predict the ethnicity of a person using their name. This classifier was trained using US Census Bureau data,³⁷ and it predicts the probabilities of a given name belonging to a person of the White, Black, Asian/Pacific Islander (API), Hispanic, American Indian/Alaskan Native (AIAN), and Mixed Race (2PRACE). A recent study by Preoțiuc-Pietro et al. [51] evaluated the performance of the `NAME-PRISM`'s [79] ethnicity classifier. They measured the Area Under the

³⁴ <https://pypi.org/project/gender-guesser/>

³⁵ <https://github.com/tue-mdse/genderComputer>

³⁶ <https://github.com/nicolasberube/Wiki-Gendersort>

³⁷ <https://www.census.gov/quickfacts/fact/note/US/RHI625221>

Table 1: Descriptive statistics of our curated dataset of 450 projects.

Metric	Min	Median	Max	Total
The number of contributors	45	164	4,033	110,336
The number of commits	1,034	4,653	151,416	4,216,386
The number of DevOps file changes	6	479	161,825	1,197,829
The number of non-DevOps file changes	2,454	26,930	813,974	29,750,595
The number of perceptibly men	24	120	3,164	104,883
The number of perceptibly women	0	10	297	9,166
The number of perceptibly White contributors	14	116	2,721	96,554
The number of perceptibly Hispanic contributors	0	4	138	3,671
The number of perceptibly API contributors	0	15	542	15,197
The number of perceptibly Black contributors	0	0	22	495
The number of perceptibly AIAN contributors	0	0	1	2

Table 2: A sample of our curated dataset.

project owner	project name	sha	datetime	file	is Dev Ops file?	author	perceived gender	perceived ethnicity
angular	angular	ef6...	2021-5-19 12:12:32.345237	.circleci /config.yml	True	Kavya	woman	API
angular	angular	c68...	2021-5-19 12:16:01.459565	docs/COMM ITTER.md	False	Pete Clay	man	White
angular	angular	c68...	2021-5-19 12:16:01.459565	docs/PUB LIC_API.md	False	Pete Clay	man	White

Receiver Operating Characteristic curve (AUC)³⁸ of NAME-PRISM with respect to the top four ethnic groups: White, API, Hispanic, and Black. Their evaluation showed that the AUC of classifying people from White, API, Hispanic, and Black were 0.719, 0.765, 0.740, and 0.681, respectively. To strengthen the validity of our study, we discard names where the ethnicity is unknown from our ethnic diversity analysis; we find only 0.03% (five) and 0.06% (59) names in which the ethnicity is unknown among DevOps and non-DevOps contributors, respectively. Furthermore, following prior studies that rely on the US Census Bureau’s classification to label developers across different countries and contexts [46], we only use contributors’ names where the perceptible ethnicity is inferred by NAME-PRISM [79] with a confidence level greater than 0.8.

The descriptive statistics of the curated dataset. Table 1 shows the descriptive statistics of our final dataset. it contains 1,197,829 DevOps file changes and 29,750,595 non-DevOps file changes made by 110,336 contributors to 450 GitHub projects, and an anonymized version of this dataset is available online along with our replication package.¹⁰ Table 2 shows a sample of data from this dataset. Some data items in this preview table are altered to protect the contributors’ identities.

(PA) Preliminary Analysis: Identifying DevOps and non-DevOps Contributors.

The goal of our study is to perform a comparison of the diversity of contributors to DevOps and non-DevOps artifacts. Thus, we conduct a preliminary analysis to

³⁸ AUC is an evaluation measurement for the classification problems, which ranges from 0–1. AUC of 0.5 represents the performance of the random classifier. The higher the AUC of a classifier than 0.5, the higher the performance of that classifier compared to the random classifier.

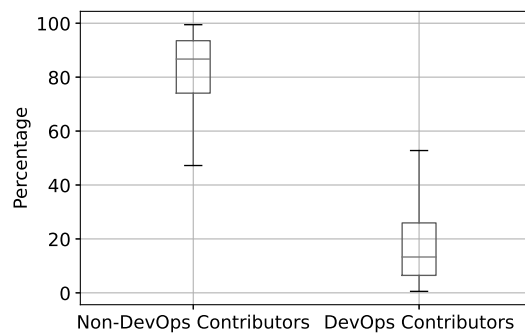


Fig. 5: Box-plots showing the percentages of contributors to non-DevOps files only and the percentages of contributors to DevOps files in the studied projects.

identify contributors to DevOps files and distinguish them from those who contribute only to non-DevOps files. To do so, for each contributor in our dataset, we check for the type of files changed in the commits made by the contributor.

Fig. 5 presents box-plots showing the distribution of contributors in a project based on their contributions to DevOps and non-DevOps artifacts. As shown in the figure, the majority of the contributors in a project (median = 86.70%) only contribute to non-DevOps files, which we refer to as *non-DevOps contributors*. However, a non-negligible percentage (median = 13.30%) of contributors made at least one commit to DevOps files. We refer to such contributors as *DevOps contributors*.

One may speculate that considering every developer who contributes to at least one DevOps file as a DevOps contributor may not be a precise definition because a DevOps contributor should make more than one DevOps file change. However, if we consider a DevOps contributor as a developer who contributes to more than one DevOps file, we need to determine the threshold number of DevOps file changes a developer should make to include in our definition of a DevOps contributor. Thus, we plot the distribution of the number of DevOps file changes made by a developer vs. the number of such developers in our dataset; Fig. 6 shows the distribution.

From the figure, we observe that using a threshold (t) of two file changes ($t = 2$) reduces the number of DevOps contributors included in our original definition from 22,580 ($t = 1$) to 16,246 ($t = 2$); using a threshold of ten file changes ($t = 10$ which is closer to a knee in the figure) reduces the number of DevOps contributors included in our original definition from 22,580 ($t = 1$) to 6,616 ($t = 10$) which is a substantial reduction of the number of DevOps contributors included in our definition; lastly, using a threshold of 286 file changes ($t = 286$ which is the knee detected by the Kneedle algorithm [61]) is extra conservative and reduces the number of DevOps contributors included in our original definition from 22,580 ($t = 1$) to 625 ($t = 286$) which is an extreme reduction of the number of DevOps contributors included in our definition.

However, setting a threshold may lead to the exclusion of developers who make valuable contributions to DevOps processes but do not meet the threshold. Therefore,

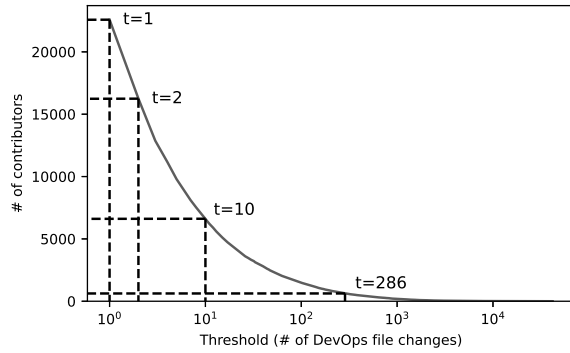


Fig. 6: Threshold plot for the number of DevOps file changes (t) made by a contributor vs. the number of contributors.

a more inclusive and realistic approach is to define a DevOps contributor as a developer who has contributed to at least one DevOps file, recognizing that even a single contribution can have a notable impact on the DevOps process and culture. Thus, for the rest of this paper, we adhere to the initial definition that we employed: contributors who made at least one commit to a DevOps file are *DevOps contributors*.³⁹

3 Independent Analysis of Ethnic and Gender Diversity

In this section, we discuss our research questions (RQs). For each RQ, we describe our approach and then present our results.

(RQ1) Does the perceptible ethnic and gender diversity of DevOps contributors differ from ethnic and gender diversity of non-DevOps contributors?

Approach. In this RQ, we quantitatively analyze the ethnic and gender diversity of DevOps and non-DevOps contributors in the studied projects. In particular, we perform two analyses. First, we compute the percentages of DevOps and non-DevOps contributors perceptible as different ethnicities and genders. Then, we statistically test the significance of the difference between the percentages of DevOps contributors and non-DevOps contributors. We use the Wilcoxon signed rank test with a 95% confidence level. We use this non-parametric test because we cannot assume our data to be normally distributed.

Second, to complement the previous analysis, we compute other common metrics including the Richness, Evenness[7, 14], Blau index (a.k.a. Diversity index/Simpson

³⁹ A sensitivity analysis with different definitions of DevOps contributors ($t=2$ and $t=10$) yielded similar results (See our online Appendices C and D ¹⁰).

index) [63], and Prevalence rankings and Diffusion score,⁴⁰ which are used to measure the diversity in a community. These metrics have been used in prior studies for similar purposes [8, 49, 72, 80]. The first metric, richness (S), measures the number of groups in a community [14]. We use the richness to measure the number of ethnicities of contributors to DevOps and non-DevOps files in a project. The higher the richness, the more diverse the community is. For example, if a project contains only contributors with perceptibly White names, the ethnic richness of that project is one. If a project contains contributors of two perceptible ethnicities (e.g., White and Asian), the richness value is two. Similarly, for gender diversity, if a project includes contributors who are perceptible as men only, the richness is one. If a project includes perceptible men as well as perceptible women, the richness is two. The second metric, evenness (E), measures the relative species abundances in a community. We use the Brillouin metric [7] to measure the evenness. Equation (1) shows the formula for computing the evenness in a community, where n_i is the number of individuals in a group i , and N is the total number of individuals in the community.

$$E = \frac{\ln(N!) - \sum_{i=1}^S \ln(n_i!)}{N} \quad (1)$$

In the context of our study, suppose a project contains 10 contributors ($N = 10$) spanning four ethnicities ($S = 4$): three perceptibly Asian contributors ($n_1 = 3$), two perceptibly Black contributors ($n_2 = 2$), one perceptibly Hispanic contributors ($n_3 = 1$), and four perceptibly White contributors ($n_4 = 4$). The evenness is computed as follows: $\frac{\ln(10!) - (\ln(3!) + \ln(2!) + \ln(1!) + \ln(4!))}{10} = 0.94$. This value shows a high level of evenness, indicating a high level of ethnic diversity of contributors. Similarly, the closer this number is to zero, the lower the evenness, thus the lower the diversity. Following the same equation, we compute the evenness for gender diversity as well.

The next metric, the Blau index, a.k.a diversity index (D) [63], computes the probability that two individuals randomly selected from a community would belong to two different groups within the community. Equation (2) shows the formula to compute the Blau index of a community. Blau index ranges from zero to one. The closer the Blau index is to one, the more diverse the community is.

$$D = 1 - \sum_{i=1}^S \left(\frac{n_i}{N}\right)^2 \quad (2)$$

In the previous example with 10 contributors, the Blau index is computed as below: $1 - \left(\left(\frac{3}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right) = 0.7$; accordingly, there is high probability of two contributors randomly selected from the project belong two different ethnicities, thus indicating a high level of ethnic diversity in the project. Similarly, we compute the Blau index for gender diversity as well.

The final diversity metric we compute in our analysis is the prevalence rankings and diffusion score.⁴⁰ The prevalence ranking in a community is the ranking of groups in the descending order of the number of individuals belonging to each group.

⁴⁰ <https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html>

From the prevalence rankings of a community, first-, second-, and third-prevalent groups could be obtained. The percentage of the individuals that are not in those first-, second-, or third-prevalent groups combined is called the diffusion score. The first-, second-, and third-prevalent ranks in the previous example are owned by perceptibly White, Asian, and Black contributors, with proportions of 40%, 30%, and 20%, respectively. Thus, the diffusion score is $100\% - (40\% + 30\% + 20\%) = 10\%$, which is equal to the percentage of least-prevalent ethnicity (Hispanic) in this example since there are only four ethnicities. For the gender diversity in the context of our study, the diffusion score will be equal to the contributor percentage from the minority gender. This is because we consider a binary classification of gender, and accordingly, there are only first- and least-prevalent groups with respect to the perceptible gender. Thus, we consider the percentage of contributors from the least-prevalent gender in a particular project/setting as the diffusion score of that case.

To test the statistical difference between diversity metrics for DevOps and non-DevOps, we first compute the metrics for each project. Then, to compare the metrics of DevOps and non-DevOps contributors, we perform the Wilcoxon signed rank test with a 95% confidence level and apply the Bonferroni correction⁴¹ to account for multiple comparisons, and our adjusted significance level (α) is 0.0023.

(RQ1-a) Perceptible Ethnic Diversity

Observation 1) Contributors to DevOps artifacts tend to have less ethnic diversity than contributors to non-DevOps artifacts.

Fig. 7 shows the distribution of the percentage of contributors to DevOps and non-DevOps artifacts in the studied projects, per ethnicity. From the figure, we observe that perceptible Whites are the majority of DevOps contributors (median = 87.70%) and non-DevOps contributors (median = 85.50%).

The second dominating ethnicity in our dataset is Asian/Pacific Islander (API). On median, the percentages of DevOps and non-DevOps contributors with perceptibly API names are 9.10% and 10.53%, respectively. We find a statistically significant difference between the percentages of DevOps and non-DevOps contributors with perceptibly API names but with a negligible effect size (Wilcoxon, $p \ll \alpha = 0.0023$, one-tailed, paired; Cliff's $|\delta| = 0.132$). The third dominating ethnicity is Hispanic. On median, the percentages of DevOps contributors with perceptibly Hispanic names are 0% and 3.25%, respectively. We find a statistically significant difference between DevOps and non-DevOps contributors with a medium effect size (Wilcoxon, $p \ll \alpha = 0.0023$, one-tailed, paired; Cliff's $|\delta| = 0.415$). The fourth dominating group is the contributors having perceptibly Black names. For those contributors, we find a statistically significant difference between the percentages of DevOps (median = 0%) and non-DevOps (median = 0.77%) contributors with a large effect size (Wilcoxon, $p \ll \alpha = 0.0023$, one-tailed, paired; Cliff's $\delta = 0.783$). Finally, we find that the contributors with perceptibly American Indian/Alaskan Native (AIAN)

⁴¹ The Bonferroni correction [5] is a well-known technique for adjusting the significance level (α) to reduce the possibility of false positives. Our entire study includes 22 statistical comparisons, and thus the adjusted α value is set to $\frac{0.05}{22} = 0.0023$.

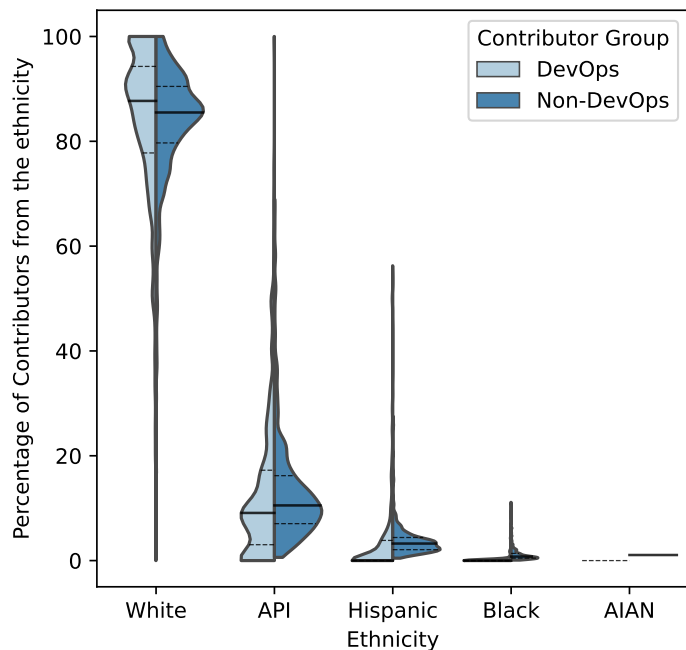


Fig. 7: Bean plots showing the distribution of percentages of DevOps and non-DevOps contributors from four perceptible ethnicities (i.e., White, API, Hispanic, and Black). The solid lines represent the median percentages, and the dotted lines represent the first and third quartiles.

Table 3: Results of statistical analysis of ethnic diversity metrics for DevOps and Non-DevOps contributors.

Diversity metric	DevOps (median)	Non-DevOps (median)	Wilcoxon, one-tailed, paired ($\alpha = 0.0023$)	Effect size (Cliff's $ \delta $)
Richness	2	3	$p = 2.81 \times 10^{-47} \ll \alpha$	0.588, considered large
Evenness	0.330	0.467	$p = 2.41 \times 10^{-31} \ll \alpha$	0.246, considered medium
Blau Index	0.219	0.255	$p = 6.95 \times 10^{-7} \ll \alpha$	0.111, considered negligible
Diffusion Score	9.091	10.157	$p = 0.006 > \alpha$	-

names are the least dominating ethnic group among DevOps contributors and non-DevOps because only two out of 450 projects we use for the analysis has contributors perceptible as AIAN. Moreover, none of the DevOps contributors in those two projects are perceptible as AIAN.

Furthermore, Table 3 presents the analysis results of ethnic diversity metrics (i.e., richness, evenness, etc.) for DevOps and non-DevOps contributors. From the table,

we observe that the median values of ethnicity richness of DevOps and non-DevOps contributors are two and three, respectively. Also, we find statistically significant differences in terms of richness of ethnic diversity between DevOps and non-DevOps contributors, with a large effect size. Similarly, the ethnic evenness of DevOps contributors is significantly less than that of non-DevOps contributors, with a medium effect size. For the Blau index, we find statistically significant differences between the ethnic diversity of DevOps and non-DevOps contributors, but the effect size is negligible. Also, note that for the diffusion score, we do not find statistical significance between DevOps and non-DevOps contributors.

(RQ1-b) Perceptible Gender Diversity

Observation 2) The perceptible gender diversity of DevOps contributors tends to be less than that of non-DevOps contributors.

Fig. 8 shows the distribution of percentages of DevOps and non-DevOps contributors who are perceptible as men and women. From the figure, we observe that the DevOps contributors who are perceptible as men are the majority of DevOps contributors (median = 93.75%) as well as non-DevOps contributors (median = 92.82%). We find a statistically significant difference between the percentages of DevOps and non-DevOps contributors who are perceptible as men, with a small effect size (Wilcoxon, $p \ll \alpha = 0.0023$, one-tailed, paired; Cliff's $|\delta| = 0.149$).

Moreover, we find that the median percentage of DevOps contributors who are perceptible as women (median = 6.25%) is less than that of non-DevOps contributors (median = 7.18%). We further find a statistically significant difference between DevOps and non-DevOps contributors who are perceptible as women, with a small effect size. This is in line with the results of the Stack Overflow survey (2022)⁴² as well. This survey shows that the percentage of DevOps developers who identify themselves as women is 2.10%, while the other developers, i.e., non-DevOps, who identify themselves as women, is 5.13%.

Furthermore, Table 4 shows the analysis results of gender diversity metrics for DevOps and non-DevOps contributors. From the table, we observe that the median value of richness in terms of gender is two for both DevOps and non-DevOps contributors. However, we find a statistically significant difference in gender richness between DevOps and non-DevOps contributors, with a small effect size. Similarly, the evenness, Blau index, and diffusion score of DevOps contributors in projects, in terms of gender diversity, are statistically less than those of non-DevOps contributors, with a small effect size. Furthermore, we find that 29.11% of the projects in our dataset did not contain any DevOps contributors perceptible as women. In contrast, only 2.00% of the projects in our dataset did not contain non-DevOps contributors perceptible as women.

Note that we exclude contributors whose perceptible gender is not determined by gender-inferring tools (Section 2 DC5). However, studying the effects of perceiving all unknown genders as either men or women can provide valuable insights into gender representation and potential biases in various contexts. Thus, we follow the

⁴² <https://insights.stackoverflow.com/survey>

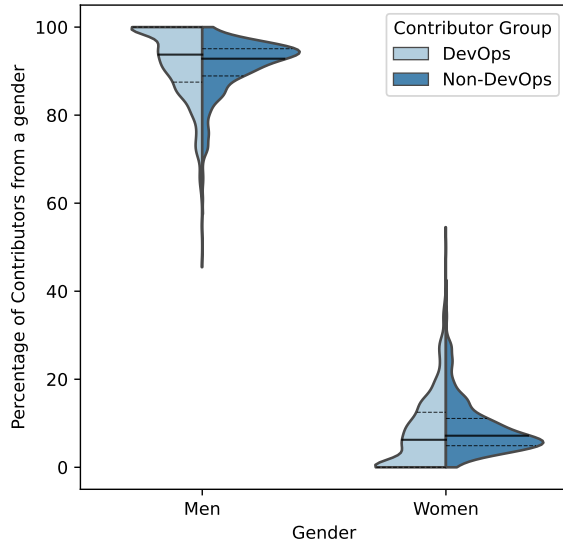


Fig. 8: Bean plots showing the distribution of percentages of contributors perceptible as men and women among DevOps contributors as well as among non-DevOps contributors. The solid lines present the median, and the dotted lines present the first and third quartiles.

Table 4: Results of statistical analysis of gender diversity metrics for DevOps and Non-DevOps contributors.

Diversity metric	DevOps (median)	Non-DevOps (median)	Wilcoxon, one-tailed, paired ($\alpha = 0.0023$)	Effect size (Cliff's $ \delta $)
Richness	2	2	$p = 1.44 \times 10^{-25} \ll \alpha$	0.271, considered small
Evenness	0.181	0.243	$p = 7.79 \times 10^{-20} \ll \alpha$	0.246, considered small
Blau Index	0.117	0.133	$p = 1.58 \times 10^{-6} \ll \alpha$	0.149, considered small
Diffusion Score	6.250	7.176	$p = 6.44 \times 10^{-5} \ll \alpha$	0.149, considered small

approach of Vasilescu et al. [70], who faced a similar issue. In particular, we investigate the impact of assuming all unknown genders as women. Since women contributors are typically underrepresented in GitHub [72], assuming all unknown genders as women would allow us to evaluate whether biases persist even when we artificially increase the representation of women contributors. From this analysis, we observe that our findings still hold: perceptible women are more underrepresented among DevOps contributors than non-DevOps contributors. A full preview of our results, assuming all contributors with unknown genders are women, is available in our online Appendix B.¹⁰

As for perceptible ethnic diversity, we observe that contributors perceptible as non-White (API, Hispanic, Black, and AIAN) are more underrepresented among DevOps contributors compared to non-DevOps contributors. With respect to perceptible gender diversity, we find contributors perceptible as women are more underrepresented among DevOps contributors compared to non-DevOps contributors. Overall, there is a statistically significant difference in the diversity metrics between DevOps and non-DevOps contributors.

(RQ2) How does the distribution of perceptible ethnic and gender diversity change as projects age?

Approach. In this RQ, we aim to examine the evolution of gender and ethnic diversity of DevOps and non-DevOps contributors in the studied projects. To do so, we partition the commits in our dataset into segments, where each segment represents a set of commits that are made during a year. Then, for each segment (year), we identify DevOps and non-DevOps contributors who contributed to the commits made during that year and analyze perceptible ethnic and gender diversity. In particular, we analyze the percentages of and compute the diversity metrics for DevOps and non-DevOps contributors in terms of ethnic (RQ2-a) and gender diversity (RQ2-b). Lastly, we compute the average growth in diversity metrics over the last ten years. Equation (3) shows how we compute the average growth, for example, in evenness, for the last ten years where $n = 10$.

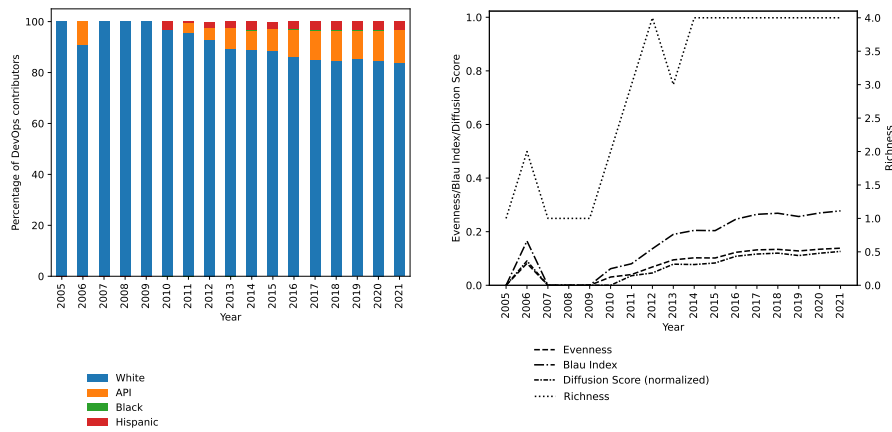
$$\text{Average growth} = \sum_{t=1}^{10} \text{evenness}_t - \text{evenness}_{t-1} \quad (3)$$

(RQ2-a) Change in Perceptible Ethnic Diversity Over Time

Observation 3) Over time, the overall perceptible ethnic diversity of DevOps and non-DevOps contributors increases. Still, the contributors perceptible as non-White (API, Hispanic, and Black) are the minority.

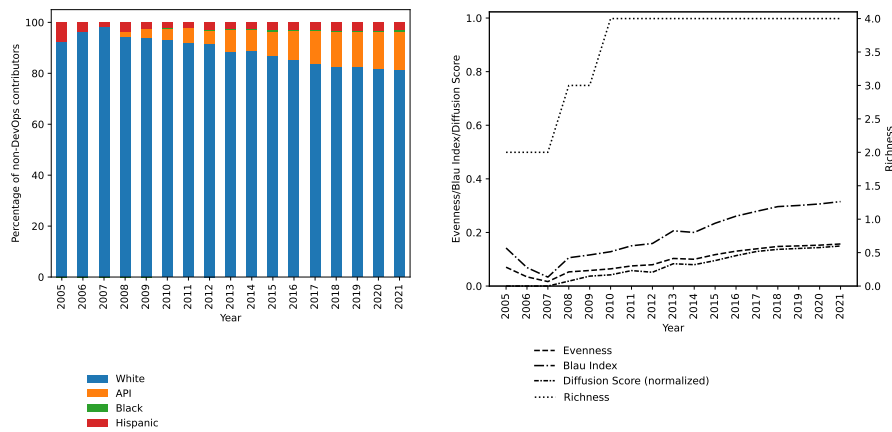
Fig. 9 shows (a) the percentage of DevOps contributors from each perceptible ethnicity per year and (b) the values of the diversity metrics per year. From Fig. 9a, we observe that the percentage of DevOps contributors who are perceptible as White generally varies between 83.98%–100%.

Also, we observe that all DevOps contributors who made commits in the studied projects before 2010 are perceptible as White (except for year 2006). Hence, the evenness metric of these years is zero (Fig. 9b). From 2010 onward, DevOps contributors with perceptibly non-White names have started to participate in the studied projects; however, the percentage remains low. For example, the percentage of DevOps contributors with API names ranges from 0%–12.63%, and the percentage of contributors with Hispanic names ranges from 0%–3.30%. This is further evident in Fig 9b. The figure shows that the diversity metrics for the ethnicity of DevOps contributors are increasing over time, yet the evenness, Blau index, and diffusion score



(a) Percentages of perceptible ethnicities of DevOps contributors over time. (b) Values of diversity metrics over time: Both the left y-axis and the right y-axis present the metrics values. Note that the diffusion score is normalized.⁴³

Fig. 9: Change in the perceptible ethnic diversity of DevOps contributors over time.



(a) Percentages of perceptible ethnicities of non-DevOps contributors over time. (b) Values of diversity metrics over time: Both the left y-axis and the right y-axis present the metrics values. Note that the diffusion score is normalized.⁴³

Fig. 10: Change in the perceptible ethnic diversity of non-DevOps contributors over time.

mostly remain constant over the last ten years. In particular, the evenness ranges from 0.068–0.138, and its average growth in evenness is 0.008 over the last ten years; the Blau index ranges from 0.136–0.278 with an average growth of 0.016; similarly,

the normalized diffusion score⁴³ ranges from 0.045–0.126, and its average growth in evenness is 0.009 over the last ten years.

The diversity of non-DevOps contributors belonging to different ethnicities follows a similar trend to that of DevOps contributors. Fig. 10 shows that diversity metrics for ethnicity of non-DevOps contributors are gradually increasing over time. We observe that the ethnic evenness, Blau index, and diffusion score for non-DevOps contributors range between 0.079–0.157, 0.158–0.315, and 0.051–0.149, respectively, over the last ten years; their average growth over time are 0.009, 0.017, and 0.010, respectively.

Note that the above observation discussing the increase of ethnic diversity over time is in line with the diversity analysis of US-based DevOps job-seekers by Zippia⁴⁴ which is a web application that is used to search US-based jobs. Zippia’s analysis revealed that the percentage of non-White developers increased from 37.16% in 2010 to 42.77% in 2021.

(RQ2-b) Change in Perceptible Gender Diversity Over Time

Observation 4) Over time, the perceptible gender diversity of DevOps and non-DevOps contributors is increasing. Still, the contributors who are perceptible as women remain underrepresented.

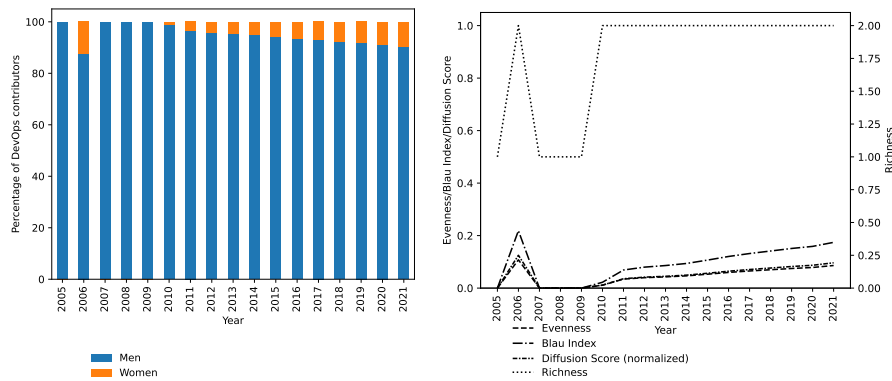
Figures 11 and 12 show (a) the percentage of perceived genders of DevOps and non-DevOps contributors and (b) the change in diversity metrics over time. We observe that the percentage of DevOps and non-DevOps contributors who are perceptible as men varies between 87.50%–100% and between 90.52%–100%, respectively. Moreover, we observe that the overall richness, evenness, Blau index, and diffusion score for DevOps and non-DevOps contributors have been increasing over time, yet the improvement is gradual. For example, the evenness for DevOps contributors over the past ten years ranges from 0.040–0.086, while it is 0.047–0.085 for non-DevOps contributors. In addition, we find that the average growth in evenness is 0.005 for DevOps contributors and 0.004 for non-DevOps contributors. We observe similar trends for the Blau index and diffusion score as well. This observation complements Prana et al.’s [50] study, which found that gender diversity in open-source projects (GHTorrent⁴⁵ dataset) has increased over time (2014–2018); however, there is still much room for improvement.

Finally, similar to the RQ1-b sensitivity analysis, for RQ2-b, we investigate the impact of assuming all contributors with unknown genders as women [70]. We find that the sensitivity analysis results do not change the observation above. A full preview of our results assuming all unknown genders are women is available in our Appendix B.¹⁰

⁴³ The diffusion score is usually a percentage, but for visualization purposes, we normalize the diffusion score by dividing the score by 100 to show the corresponding proportion.

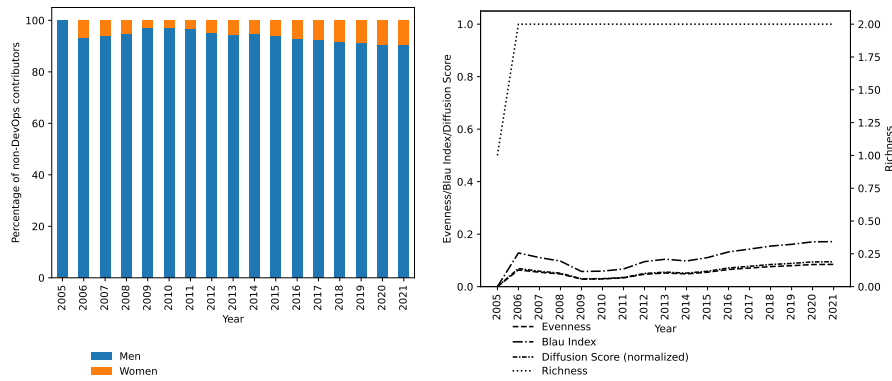
⁴⁴ <https://www.zippia.com/devops-engineer-jobs/demographics/>

⁴⁵ <http://ww38.ghtorrent.org>



(a) Percentages of perceptible gender of DevOps contributors over time. (b) Values of diversity metrics over time: Both the left y-axis and the right y-axis present the metrics values. Note that the diffusion score is normalized.⁴³

Fig. 11: Change in the perceptible gender diversity of DevOps contributors over time.



(a) Percentages of perceptible gender of non-DevOps contributors over time. (b) Values of diversity metrics over time: Both the left y-axis and the right y-axis present the metrics values. Note that the diffusion score is normalized.⁴³

Fig. 12: Change in the perceptible gender diversity of non-DevOps contributors over time.

Although the perceptible ethnic diversity of contributors is increasing over time, DevOps and non-DevOps contributors with perceptibly non-White names are greatly underrepresented. Similarly, the perceptible gender diversity of contributors is gradually increasing over time. Still, perceptible women remain substantially underrepresented.

4 Intersectional Analysis of Perceptible Ethnic and Gender Diversity

Our study thus far has been performed to independently analyze the perceptible gender and ethnic diversity of DevOps and non-DevOps contributors. However, it is unclear how diverse DevOps and non-DevOps contributors are when considering the intersection of minority groups of perceptible ethnicity and gender (e.g., perceptibly women contributors with perceptibly non-White names). In fact, prior studies [2, 28, 43, 50, 57, 68] have shown that individuals who belong to the intersection of two minority groups face specific challenges in the fields of Science, Technology, Engineering, and Mathematics (STEM). Indeed, contributors who belong to two or more underrepresented categories tend to receive fewer opportunities [28].

Hence, to shed light on the intersectionality of gender and ethnic diversity of DevOps and non-DevOps contributors, we examine the participation of contributors with perceptibly non-White and women’s names. We first compute the percentage of DevOps and non-DevOps contributors perceptible as women of different perceived ethnicities. Then, we test the significance of the difference between such DevOps and non-DevOps contributors in the studied projects using the Wilcoxon signed rank test with a 95% confidence level (with Bonferroni correction; $\alpha = 0.0023$). Lastly, to examine the trend of intersectionality over time, we compute the diversity metrics (e.g., richness and evenness) for DevOps and non-DevOps contributors per year.

Observation 5) Among perceptibly women DevOps and non-DevOps contributors, perceptibly White contributors are the majority.

Table 5 shows the percentage of perceptible ethnicities of perceptibly women DevOps and non-DevOps contributors (since none of the DevOps contributors in our dataset are perceptible as AIAN, we do not consider the AIAN in this intersectional analysis). From the table, we observe that contributors with perceptibly White names are the majority among perceptibly women DevOps contributors (median = 100%) and non-DevOps contributors (median = 70%) in a project. The second ethnic majority among perceptibly women contributors is API; the median percentages of DevOps and non-DevOps contributors are 0% and 25%, respectively. Perceptibly women contributors with perceptibly Hispanic and Black names are the least included among DevOps and non-DevOps contributors (median = 0%).

From the table, we also observe statistically significant differences in the perceived ethnic diversity between DevOps and non-DevOps contributors with perceptibly women names. In particular, the contributors who are perceptible as White

Table 5: Results of the statistical analysis of the percentages of the perceptible ethnicities of perceptibly women DevOps and non-DevOps contributors.

Perceptible Group	DevOps (median)	Non-DevOps (median)	Wilcoxon, one-tailed, paired ($\alpha = 0.0023$)	Effect size (Cliff’s $ \delta $)
White women	100%	70%	$p = 5.85 \times 10^{-5} \ll \alpha$	0.221, considered small
API women	0%	25%	$p = 3.48 \times 10^{-4} < \alpha$	0.219, considered small
Hispanic women	0%	0%	$p = 3.07 \times 10^{-6} \ll \alpha$	0.193, considered small
Black women	0%	0%	$p = 0.18 > \alpha$	-

women are more represented in DevOps compared to non-DevOps, with a small effect size (Wilcoxon, $p \ll \alpha = 0.0023$, one-tailed, paired; Cliff's $|\delta| = 0.221$). With respect to contributors who belong to the intersection of two minorities, perceptibly API women and Hispanic women contributors are more underrepresented among DevOps contributors compared to non-DevOps contributors, with a small effect size. Lastly, we do not observe a significant difference between the DevOps and non-DevOps contributors who are perceptible as black women. However, we observe that a substantial proportion of projects do not contain such contributors. In particular, 34% of the projects do not have perceptibly black women among DevOps contributors, while 4% of the projects do not have any perceptibly black women among non-DevOps contributors.

Observation 6) Over time, the perceptible ethnic diversity of perceptibly women DevOps and non-DevOps contributors is increasing. Perceptibly women contributors with perceptibly White names remain the majority for both DevOps and non-DevOps contributors.

Fig. 13 presents (a) the percentages of different perceptible ethnicities of perceptibly women DevOps contributors per year, and (b) the values of the corresponding diversity metrics for perceptibly women DevOps contributors per year. From Fig. 13a, we observe that the number of ethnicity groups of DevOps contributors with perceptibly women names is increasing. For example, the percentage of perceptibly API women contributors has increased from 0% in 2010 to 22.9% in 2021, excluding the exception of 2006). This is further evident in Fig. 13b, as it shows an overall increase in diversity metrics with respect to the perceived ethnic diversity of perceptibly women DevOps contributors. For example, the ethnic evenness of perceptibly women DevOps contributors increased from zero in 2010 to 0.218 in 2021. Still, we observe that the improvement in evenness over the last few years is not substantial; the average growth in evenness over the last ten years is 0.007. We observe similar trends for the Blau index and diffusion score as well.

For a subset of authors (7.4%), ethnic and gender-inferring tools could not identify both the perceptible ethnicity and the gender with the expected confidence level. Thus, for years 2005, 2007, 2008, and 2009 in Fig. 13a, we do not observe perceptibly women DevOps contributors with a perceptible ethnicity confidence level greater than 0.8. For 2006, we find an exception where only one perceptibly women DevOps contributor is available with a perceptible ethnicity confidence level greater than 0.8 in our dataset.

With respect to non-DevOps contributors, Fig. 14 shows that perceptibly women contributors of different ethnicities follow a similar trend to DevOps contributors. In particular, the richness is increasing, yet the other metrics have remained almost constant in the last ten years; for example, the evenness ranges between 0.165–0.229, and the average growth is only 0.005.

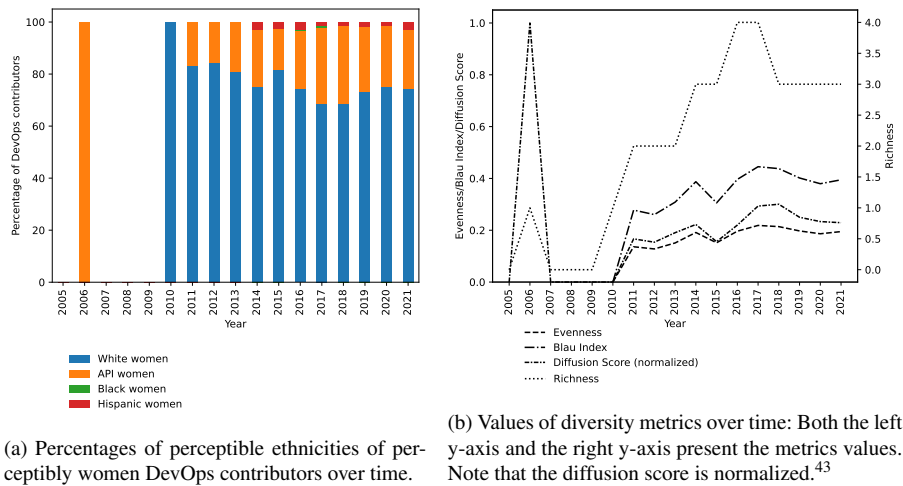


Fig. 13: Change in the perceptible ethnic diversity of the perceptibly women DevOps contributors.

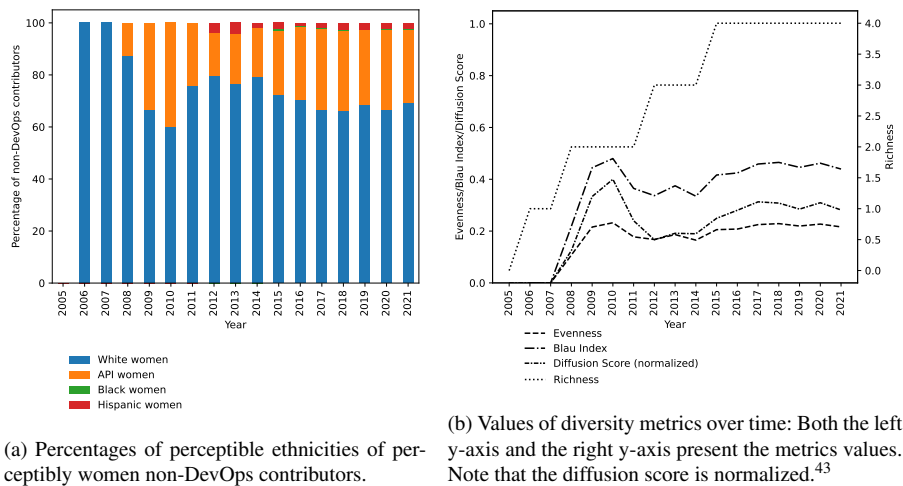


Fig. 14: Change in the perceptible ethnic diversity of the perceptibly women non-DevOps contributors.

Contributors who are perceptible as White women are the majority of perceptible women contributors, followed by contributors with perceptibly API, Hispanic, and Black names. We find statistical differences between the participation of perceptible women in DevOps and non-DevOps (e.g., contributors who are perceptibly API and Hispanic women are more underrepresented in DevOps compared to non-DevOps). Additionally, it is worth noting that perceptibly black women are absent among DevOps contributors in 34% of the projects, while among non-DevOps contributors, this is only 4%. Over time, the richness of perceptible ethnic diversity of perceptibly women contributors tends to increase, yet the diversity metrics remain low.

5 Threats to Validity

In this section, we describe the threats to the validity of our study.

5.1 Construct Validity

We use the NAME-PRISM [79] tool to infer the perceptible ethnicity of contributors. Using NAME-PRISM to infer ethnicity has three main limitations. First, the ethnicity of an individual is a complex and multifaceted social construct that is not always easily identifiable [19], and the tool may misassign ethnicities to contributors. For example, contributors with Brazilian names but are not Hispanic (because their ancestors do not speak Spanish) may be perceived as Hispanic by the NAME-PRISM. Mixed-race contributors also may be perceived as belonging to a particular ethnicity, which may not necessarily align with the ethnicity that they identify with personally. Also, some Black individuals may have names perceived as White [22], and using the perceived ethnicity may exclude individuals who identify as Black. However, note that in open-source communities, such as GitHub, one would only perceive the ethnicity of the other unless and otherwise it is revealed. Thus, our goal is to study the perceived diversity instead of the actual diversity. Moreover, as suggested by prior work [46], we only use contributors' names in which the perceptible ethnicity is inferred by the tool with a confidence level greater than 0.8 for better validity of our results.

The second limitation of using NAME-PRISM is that the US government officially categorizes people with origins in Lebanon, Iran, Egypt, and other countries in the Middle East and North Africa (MENA) region as White.⁴⁶ According to Maghbouleh et al. [42], both non-MENA Whites and MENA individuals consider MENA-related traits as MENA rather than White. In addition, treating individuals from the MENA region as Whites does not necessarily correspond to the discrimination-related experiences of many of these people.⁴⁷ Because of this, considering individuals from the MENA region as White as per US Census Bureau may not accurately reflect the perceptible diversity of the DevOps and non-DevOps contributors we study. Therefore, to check the validity of our conclusions, we recompute our results by considering the perceptibly White contributors with perceptible MENA-related traits as a separate ethnic group. However, the contributor names available in our current dataset do not include the necessary information (e.g., nationality) to check if a contributor is associated with a MENA-related trait or not. Thus, we use NAME-PRISM's perceptible nationality-inferring API (which has an F-score of 0.795). We find our conclusions remain unchanged. A full preview of our results corresponding to this new distinction of contributors perceptible as from MENA countries is available in our Appendix E.¹⁰

⁴⁶ <https://www.documentcloud.org/documents/21218285-2020-census-state-pl-94-171-redistricting-summary-file-technical-documentationdocument/p232/a2082431>

⁴⁷ <https://www.npr.org/2022/02/17/1079181478/us-census-middle-eastern-white-north-african-mena>

The third limitation of using the NAME-PRISM is that the results obtained with respect to US Census Bureau’s classification may not be completely generalizable for contributors who are based outside the US [2, 59]. To reflect on the validity of our study, we perform a new analysis to understand the impact of using NAME-PRISM for both US-based and non-US-based contributors. To categorize contributors into “US-based” and “Outside US,” we attempt to use the GitHub API to obtain the geographical locations, specifically the countries, of the contributors within our dataset. Because our current dataset does not include GitHub usernames, which are exclusively needed to retrieve locations via GitHub API,⁴⁸ we manually check for GitHub profiles of those contributors for locations. Note that NAME-PRISM’s perceptible nationality-inferring API that we used in our previous analysis on contributors perceptibly associated with MENA-related traits does not infer locations but rather nationalities; for example, it does not make the distinction between perceptible Whites from the US vs. the perceptible Whites from other countries. For our manual check, we obtain a significant sample of 500⁴⁹ contributors by searching on GitHub, and we check if there is a substantial difference in the observations we made with respect to the two distinct groups based on their locations: “US-based” and “Outside US.” We do not observe substantial changes in the observations concerning the two groups. For example, we find that, regardless of the location, perceptibly White contributors are the majority; contributors from other perceptible ethnicities are generally underrepresented. This is consistent with the observations we made in Section 3. A full preview of our results corresponding to this new analysis is available in our Appendix J.¹⁰ Furthermore, to facilitate future studies on diversity, we encourage researchers to train more sophisticated ethnicity-inferring tools with datasets spanning many countries and diverse populations.

We use gender-inferring tools (e.g. Gender-guesser) to infer the perceptible gender of contributors. Using such tools to infer gender has two main limitations. First, gender-inferring tools may not recognize the perceptible gender correctly as discussed in Section 2 DC4 (e.g., rate of unrecognized names in Gender-guesser was 20.12% [60]). We tackle this problem by combining the outcomes from three gender-inferring tools to take advantage of the strengths of all the tools as discussed in Section 2. We use the most effective combination, which we obtain by testing different combinations of the tools against the labelled dataset provided by Santamaría et al. [60]). The selected combination approach reduces the rate of unrecognized names to 5.15%. Then, we removed the names with unrecognized genders from the analyses on perceived gender diversity for better accuracy.

The second limitation of using gender-inferring tools is that gender-inferring tools have different classifications of gender given a name. To ensure consistency across all the tools used in this study, we follow the definition of gender as a binary classification (either men or women) as used in prior studies by Vasilescu et al. [69, 70, 72]. As mentioned in Section 2 (DC5), we use three gender inferring

⁴⁸ <https://docs.github.com/en/free-pro-team@latest/rest/users/users?apiVersion=2022-11-28#get-a-user>

⁴⁹ The sample size of 500 individuals exceeds the minimum recommended threshold (383) for achieving a 95% confidence level with a 5% margin of error when making inferences about the entire contributor population within our dataset.

tools: Wiki-Gendersort, Gender-guesser, and GenderComputer. The first tool, the Wiki-Gendersort, classifies a name into one of the three gender groups: male, female, and unisex. We consider female category as women and male category as men; we consider the remaining category, unisex, to be unknown because we limit our discussion to a binary classification of gender [69, 70, 72]. The next tool, the Gender-guesser, classifies a name into one of the five categories: female, mostly female, androgynous, mostly male, and male. We consider female and mostly female categories as women; similarly, we consider male and mostly male categories as men. We consider the Gender-guesser's remaining category (androgynous) as unknown, as similar to Wiki-Gendersort's unisex category. The third tool, the GenderComputer, classifies gender into two main categories: male and female.⁵⁰ Although we are aware that gender is a more complex issue than a binary variable We consider the female category inferred from GenderComputer to be women, and the male category to be men. Although we are aware that gender is a more complex issue than a binary variable [60], we believe that following the terminologies used in prior work mitigates the risk of misrepresenting genders. Future researchers may consider training more sophisticated gender-inferring tools with much larger datasets taking the diverse representations of gender into consideration.

We use Vasilescu et al.'s GitHub-alias-merging approach [73] to track name changes of users, as discussed in Section 2 (DC4). If a user changes the name to a completely different one but keeps using the same email address, we are able to unify the identities of such cases. However, if the user makes substantial changes to both their name and email address, we may miss such cases. In fact, it is infeasible to recognize such cases by human contributors as well since no ground truth data is available for users. Nonetheless, we only discuss the perceived diversity in this paper, and missing such extreme cases would not substantially impact our study's conclusions.

Lastly, in our study, we use four measures of diversity (richness [14], evenness [7], Blau index [63], and prevalence rankings, and diffusion score⁴⁰). Other indices, such as Gini index [27] and Theil index [66], are not applicable to our study because they are not comparable across different teams when teams have different numbers of ethnicities or genders, as in our study.

5.2 Internal Validity

To identify DevOps files, we rely on filename and keyword conventions. Despite our best effort, our automated classification script may still misclassify files. On the other hand, our initial results from a validation set (of 400 files) show that the agreement between a human labeller and the automated classification script is near-perfect (Cohen's Kappa = 0.82).

Our preliminary analysis of the dataset shows that some DevOps files are automatically generated and hence, are not valid targets for our analyses. We filter out

⁵⁰ Note that we cannot claim Gender-guesser's androgynous category and Wiki-Gendersort's unisex category to have the same meaning without evidence.

such files using keyword matching. Initial results from a validation set of 400 randomly selected files that are classified as DevOps show that the agreement between a human labeller and the automated classification script is near-perfect (i.e., Cohen’s Kappa of 0.81).

We considered DevOps contributors as the ones who made at least one change to DevOps files. A threshold of one DevOps file change may introduce selection bias to our definition of DevOps contributors. As discussed in Section 2, we recompute our results with two more thresholds: two file changes and ten file changes. We find our results still hold, indicating that our choice of the threshold does not include a substantial selection bias.

5.3 External Validity

We begin with a dataset of projects using CircleCI, and carefully select a meaningful sample of projects that are worthy of study. To do so, we filter out the projects that are immature and have little development history by applying filtration criteria used by prior studies [38]. This selection results in a dataset of 450 projects that adopt DevOps tools and technologies. Our dataset might be considered small when it is compared to the whole population of GitHub projects. To check the validity of our analyses beyond the projects that adopt CircleCI, we recompute the results of our research questions using another dataset that contains commits of projects that adopt GitHub Actions.⁵¹ We find that the overall conclusions of the study remain unchanged for this new dataset of projects that adopt GitHub Actions. To access a complete preview of this analysis and its corresponding figures, we direct readers to our Appendix F.¹⁰

6 Related work

In this section, we discuss the most work related to ours. In particular, we describe the prior work that discusses diversity in the context of software engineering.

A plethora of prior work focused on analyzing the diversity of project teams [9, 36, 71, 74]. For example, Vasilescu et al. [74] analyzed diversity data for 23,493 GitHub projects. They found that 75.3% of the studied projects have no gender diversity at all. On the other hand, Canedo et al. [9] surveyed 34 manually gender-confirmed women core developers and found that most of them (65.7%) have never experienced gender discrimination. Canedo et al. [9] performed another analysis of commits, and they found that the total number of commits and the frequency of commits made by a woman developer are not significantly different from those made by a man developer.⁵² However, Canedo et al. analyzed commits and commit messages of contributors in 711 GitHub projects, and it revealed that women developers tend to

⁵¹ <https://github.com/features/actions>

⁵² Inspired by Canedo et al. [9], we perform a similar analysis for DevOps and non-DevOps contributors to see if there is any difference in the total number of commits and the frequency of commits made by a perceptibly woman contributor compared to a man contributor. Our results do not reveal any changes, and that complements Canedo et al.’s results. A full preview of results is available in our online Appendix H.¹⁰

contribute more to re-engineering tasks than men developers. Vasilescu et al. [71] surveyed GitHub developers to identify their perceptions of diversity on GitHub. They found that 48.24% of the developers are aware of the gender of their fellow developers, while it is 29.94% for ethnicity. Izquierdo et al. [36] analyzed the diversity of OpenStack, a well-known open-source project involving thousands of contributors. They found that women make up 10% of all OpenStack contributors.

Other work investigated how the demographic characteristics of contributors can influence the acceptance of Pull Requests (PRs) [23, 35, 45, 46, 53, 54, 65]. Terrell et al. [65] found that, among external contributors, women whose gender is identifiable have a lower PR acceptance rate. Rastogi et al. [53, 54] also found that PRs are more likely to get accepted when both the submitter and the integrator are from the same geographical location. Moreover, Nadri et al. [45] found contributors with non-White races have a tendency to receive rejections without rationale. Later, Nadri et al. [46] found that PRs from perceptibly White developers have a higher acceptance chance, and perceptibly non-White contributors are more likely to get their PRs accepted if the integrator is also from the same race. Furtado et al. [23] found that contributors from countries with low development in health, education, and income submit fewer PRs but receive the most rejections. Iyer et al. [35] found that personality traits and social factors are significant and comparable to technical factors that influence the likelihood of PR acceptance.

Other research work focused on the relationship between diversity and team productivity [6, 10, 32, 65, 72]. Bosu et al. [6] studied ten popular open-source projects hosted on GitHub. They found that despite no project indicating any significant productivity differences between men and women developers, women developers tend to receive slower feedback during code reviews, and are underrepresented in leadership roles. Catolino et al. [10] studied the gender imbalance of developers in projects and its effect on community smells (e.g., deliberate cessation of communications and defiant community members). They found that women were more restrained in the discussion forums, regardless of how senior or productive they were. Terrell et al. [65] studied 4,500 GitHub contributors, and found that gender diversity is a significant and positive predictor of productivity. Vasilescu et al. [72] and Ortu et al. [48] have also studied gender diversity in GitHub projects, and found that gender diversity has a positive effect on team productivity. However, a few studies found contradictory results as well [26, 32]. For example, Horwitz et al. [32] found that bio-demographic (e.g., gender and ethnicity) diversity has no relationship with team performance.

Prior studies have also investigated the diversity of developers participating in online socio-technical platforms [47, 69]. Vasilescu et al. [69] studied the gender representation in Stack Overflow (a Q&A platform for developers). They found that men represent the vast majority of contributors to Stack Overflow. Also, they found that men earn more reputation and are more engaged than women. Oliveira et al. [47] studied the design of Stack Overflow and contrasted their findings with the views of participants of various cultural backgrounds. They found that the design of Stack Overflow follows individualist values, such as productivity and reputation scores, which can discourage the engagement of those with collectivist values.

Recent surveys set out to understand the trend of the research area concerned with perceived diversity in Software Engineering [56, 67]. Pérez et al. [56] highlighted that

researchers have been exploring the gender bias problems in software engineering more than presenting the solutions to mitigate these problems. Trinkenreich et al. [67] analyzed the literature to identify strategies proposed to mitigate the challenges faced by women in open-source projects, such as de-stereotyping the open-source project contributors, training mentors to assist women, creating and enforcing a code of conduct, and placing women in leadership positions.

Several studies [2, 28, 43, 50, 57, 68] indicated that individuals who identify themselves as in the intersection of two or more minority groups encounter specific obstacles in technical fields. For example, Black women tend to have less exposure to Software Engineering [57] and face systematic impediments to upward career mobility [28]. Prana et al. [50] surveyed 122 developers worldwide and measured gender diversity with respect to different regions using the Blau index [63]. Prana et al. found that gender diversity is low worldwide, but there is a perceived distinction in diversity across regions that is not statistically substantial; for example, the gender diversity of respondents from Africa is lower than that among respondents from America. Our intersectional analysis (of ethnicity and gender) adds a complementary perspective to these existing studies, since we analyze the representation of developers with respect to their contributions: i.e., contributors to DevOps artifacts and non-DevOps artifacts. Our findings show that the lack of perceived diversity among DevOps contributors is amplified when considering the intersection of ethnicity and gender”

Several surveys and/or analyses by online communities have also raised concerns about the diversity of DevOps developers: yearly surveys of Stack Overflow and the US-based analysis from “Zippia.”⁵³ Both of the above show that men are the majority among DevOps developers. For example, the Stack Overflow survey (2022)⁵⁴ reported that the percentage of developers who identified themselves as DevOps developers is 10.06%, and of them, the vast majority are men (94.37%). A study by Wurzelova et al. [78] also analyzed the dataset of the Stack Overflow survey (2018).⁵⁵ They found that 27.41% of the women who report to contribute to open source are DevOps specialists. In the context of our study, we find that 19.89% of the perceptibly women contributors in open source are also contributing to DevOps artifacts. While the Stack Overflow-related analyses and Zippia’s analysis provide insights into the overall diversity of the DevOps community, our study complements them with a contribution-based analysis of perceptible ethnic and gender diversity at the project level and over time. Moreover, we ground our analysis in comparison with non-DevOps contributors in the same set of studied projects.

7 Conclusion

In this paper, we investigate the perceived gender and ethnic diversity of DevOps and non-DevOps contributors to open-source projects by quantitatively analyzing 4,207,735 commits made by 110,336 contributors to 450 open-source GitHub projects. Below, we discuss the implications of our findings.

⁵³ <https://www.zippia.com/devops-engineer-jobs/demographics/>

⁵⁴ <https://survey.stackoverflow.co/2022/>

⁵⁵ <https://survey.stackoverflow.co/2018/>

The lack of perceived diversity in the project subteams (i.e., DevOps) is even more prominent than for teams overall. Our findings indicate disparities in the perceived ethnic and gender representation among DevOps and non-DevOps contributors in the GitHub community. In particular, our observations (1) and (2) show that the group of contributors who contribute to DevOps artifacts tend to have less perceptible diversity than contributors to non-DevOps artifacts. Prior studies [39, 41] have discussed several challenges of adopting DevOps, such as difficulties in choosing appropriate tools from a diverse set of tools, which may be contributing to the lack of perceptible diversity. Potential reasons for the lack of diversity in DevOps have been discussed in various online reports as well. For instance, the report “Inspiring Women to Join the DevOps Movement”⁵⁶ discussed a lack of awareness among women developers about the availability of career paths such as DevOps. Several reports^{57 58} also discussed the stereotypical perception that DevOps is man-dominated, which may discourage women from pursuing careers in this area. Concerns were raised about the gender pay gap^{59 60} in DevOps, which may also contribute to the lack of diversity. We encourage future studies to direct efforts towards further understanding why DevOps contributions are less attractive to perceptibly non-Whites and women developers.

The perceived diversity of DevOps and non-DevOps contributors is slowly increasing over time, but there is still room for improvement. Observations (3) and (4) show that perceptible diversity is increasing over time, but the improvement is not substantial. To bridge this gap, one particular effort is the “Women in DevOps” platform,⁶¹ which was established in 2017 specifically to address the issue of gender imbalance in the DevOps industry. Furthermore, to improve diversity in Software teams in general, several studies [46, 50, 76] have suggested strategies. For example, Wang et al. [76] have proposed to design a series of carefully crafted and empirically tested training courses that aim to reduce gender bias in both educational institutions and software development organizations. Prana et al. [50] emphasized the fact that current automatic tools, such as bug assignment tools [37, 40], make recommendations based on the similarity between developers (homophily), restricting the promotion of diversity. That said, to the best of our knowledge, it is yet to investigate whether projects get the benefit of implementing such strategies. Future work could examine the impact of various strategies, such as mentorship programs [50, 76] and code of conduct amendments [20, 50], shedding light on specific interventions required to promote diversity and inclusion for developers contributing to different project activities (e.g., DevOps and non-DevOps).

⁵⁶ <https://www.pagerduty.com/blog/inspiring-women-to-join-the-devops-movement/>

⁵⁷ <https://www.cloudbees.com/blog/women-devops-tracy-ragan>

⁵⁸ <https://peapletalent.com/the-gender-gap-in-technology>

⁵⁹ <https://www.puppet.com/press/press-releases/puppets-seventh-annual-devops-salary-report>

⁶⁰ <https://techmonitor.ai/leadership/workforce/devops-gender-pay-gap-critics-calling-change>

⁶¹ <https://www.womenindevops.com/>

The lack of perceived diversity is amplified when considering the intersection of ethnicity and gender of DevOps and non-DevOps contributors. Much prior work (e.g., [6, 45, 46, 72]) that studied diversity has mainly considered independent analyses of ethnic and gender diversity. Considering the intersection of perceived ethnic and gender diversity is important to provide a richer understanding of the problems that individuals in this intersection encounter. For example, our observations (5) and (6) show that DevOps and non-DevOps contributors perceptible as Hispanic women and Black women are greatly underrepresented (median = 0%), while DevOps contributors perceptible as API women and Hispanic women tend to be more underrepresented than non-DevOps contributors. We believe that the challenges faced by those who are perceptible at the intersection of a minority ethnicity and minority gender may not be treated with a single overarching solution targeted towards one identity factor. Thus, we encourage future studies to explore those challenges and amplify the voices of developers situated at the intersection.

Acknowledgements

We would like to thank Muhammad Arsalan Khan for their help in collecting related papers for our study. Next, we would also like to thank Ajiromola Kola-Olawuyi for sharing their experiences of using various DevOps tools and for helping us in creating the list of DevOps tools for the analysis.

Conflict of Interest

The authors declared that they have no conflict of interest.

References

1. Periodic table of devops. URL <https://digital.ai/periodic-table-of-devops-tools>
2. Albusays, K., Bjorn, P., Dabbish, L., Ford, D., Murphy-Hill, E., Serebrenik, A., Storey, M.A.: The diversity crisis in software development. *Software* **38** (2021)
3. Bang, S.K., Chung, S., Choh, Y., Dupuis, M.: A grounded theory analysis of modern web applications: knowledge, skills, and abilities for devops. In: Proceedings of the 2nd annual conference on Research in information technology (2013)
4. Bass, L., Weber, I., Zhu, L.: DevOps: A software architect's perspective (2015)
5. Bonferroni, C.: Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze **8** (1936)
6. Bosu, A., Sultana, K.Z.: Diversity and inclusion in open source software (oss) projects: Where do we stand? In: Proceedings of the International Symposium on Empirical Software Engineering and Measurement (2019)

7. Brillouin, L.: Maxwell's demon cannot operate: Information and entropy. i. *Journal of Applied Physics* **22** (1951)
8. Buch, T., Meister, M., Niebuhr, A.: Ethnic diversity and segregation in german cities. *Cities* **115** (2021)
9. Canedo, E.D., Bonifácio, R., Okimoto, M.V., Serebrenik, A., Pinto, G., Monteiro, E.: Work practices and perceptions from women core developers in oss communities. In: *Proceedings of the 14th International Symposium on Empirical Software Engineering and Measurement* (2020)
10. Catolino, G., Palomba, F., Tamburri, D.A., Serebrenik, A., Ferrucci, F.: Gender diversity and women in software teams: How do they affect community smells? In: *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Society* (2019)
11. Cohen, J.: Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* **70** (1968)
12. Cukier, D.: Devops patterns to scale web applications using cloud services. In: *Proceedings of the companion publication for conference on Systems, programming, & applications: software for humanity* (2013)
13. Davidson, J.L., Naik, R., Mannan, U.A., Azarbakht, A., Jensen, C.: On older adults in free/open source software: reflections of contributors and community leaders. In: *Proceedings of the Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (2014)
14. DeJong, T.M.: A comparison of three diversity indices based on their components of richness and evenness. *Oikos* (1975)
15. Earley, C.P., Mosakowski, E.: Creating hybrid team cultures: An empirical test of transnational team functioning. *Academy of Management journal* **43** (2000)
16. Ebert, C., Gallardo, G., Hernantes, J., Serrano, N.: Devops. *Software* (2016)
17. Else, H., Perkel, J.M., et al.: The giant plan to track diversity in research journals. *Nature* **602** (2022)
18. Erich, F.M., Amrit, C., Daneva, M.: A qualitative study of devops usage in practice. *Journal of software: Evolution and Process* **29** (2017)
19. Fearon, J.D., Laitin, D.D.: Violence and the social construction of ethnic identity. *International organization* **54** (2000)
20. Ford, D., Milewicz, R., Serebrenik, A.: How remote work can foster a more inclusive environment for transgender developers. In: *Proceedings of the 2nd International Workshop on Gender Equality in Software Engineering* (2019)
21. Ford, D., Smith, J., Guo, P.J., Parnin, C.: Paradise unplugged: Identifying barriers for female participation on stack overflow. In: *Proceedings of the 24th SIGSOFT International symposium on foundations of software engineering* (2016)
22. Fryer Jr, R.G., Levitt, S.D.: The causes and consequences of distinctively black names. *The Quarterly Journal of Economics* **119** (2004)
23. Furtado, L.B., Cartaxo, B., Treude, C., Pinto, G.: How successful are open source contributions from countries with different levels of human development? *Software* **38** (2020)
24. Galinsky, A.D., Todd, A.R., Homan, A.C., Phillips, K.W., Apfelbaum, E.P., Sasaki, S.J., Richeson, J.A., Olayon, J.B., Maddux, W.W.: Maximizing the gains and minimizing the pains of diversity: A policy perspective. *Perspectives on*

- Psychological Science **10** (2015)
25. Gallaba, K., Lamothe, M., McIntosh, S.: Lessons from eight years of operational data from a continuous integration service: An exploratory case study of circleci. In: Proceedings of the 44th International Conference on Software Engineering (2022)
 26. Gila, A.R., Jaafa, J., Omar, M., Tunio, M.Z.: Impact of personality and gender diversity on software development teams' performance. In: Proceedings of the International Conference on Computer, Communications, and Control Technology (2014)
 27. Gini, C.: On the measure of concentration with special reference to income and statistics. Colorado College Publication, General Series **208** (1936)
 28. Green, H.: Disparity in discrimination: A study on the experience of minority women in the workplace (2017)
 29. Gren, L.: On gender, ethnicity, and culture in empirical software engineering research. In: Proceedings of the 11th International Workshop on Cooperative and Human Aspects of Software Engineering (2018)
 30. Griffin, D., Li, K., Xu, T.: Board gender diversity and corporate innovation: International evidence. Journal of Financial and Quantitative Analysis **56** (2021)
 31. Gușeală, L.G., Bratu, D.V., Moraru, S.A.: Devops transformation for multi-cloud iot applications. In: Proceedings of the International Conference on Sensing and Instrumentation in IoT Era (2019)
 32. Horwitz, S.K., Horwitz, I.B.: The effects of team diversity on team outcomes: A meta-analytic review of team demography. Journal of management **33** (2007)
 33. Hui, J.S., Farnham, S.D.: Designing for inclusion: Supporting gender diversity in independent innovation teams. In: Proceedings of the 2016 ACM International Conference on Supporting Group Work (2016)
 34. Humble, J., Molesky, J.: Why enterprises must adopt devops to enable continuous delivery. Cutter IT Journal **24** (2011)
 35. Iyer, R.N., Yun, S.A., Nagappan, M., Hoey, J.: Effects of personality traits on pull request acceptance. Transactions on Software Engineering **47** (2019)
 36. Izquierdo, D., Huesman, N., Serebrenik, A., Robles, G.: Openstack gender diversity report. Software **36** (2018)
 37. Jonsson, L., Borg, M., Broman, D., Sandahl, K., Eldh, S., Runeson, P.: Automated bug assignment: Ensemble-based machine learning in large scale industrial contexts. Empirical Software Engineering **21** (2016)
 38. Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D.M., Damian, D.: An in-depth study of the promises and perils of mining github. Empirical Software Engineering **21** (2016)
 39. Khan, M.S., Khan, A.W., Khan, F., Khan, M.A., Whangbo, T.K.: Critical challenges to adopt devops culture in software organizations: a systematic review (2022)
 40. Lee, S., Wu, R., Cheung, S.C., Kang, S.: Automatic detection and update suggestion for outdated api names in documentation. Transactions on Software Engineering **47** (2019)
 41. Leite, L., Rocha, C., Kon, F., Milojevic, D., Meirelles, P.: A survey of devops concepts and challenges. Computing Surveys **52** (2019)

42. Maghbouleh, N., Schachter, A., Flores, R.D.: Middle eastern and north african americans may not be perceived, nor perceive themselves, to be white. *Proceedings of the National Academy of Sciences* **119** (2022)
43. Modi, K., Schoenberg, J., Salmond, K.: *Generation stem: What girls say about science, technology, engineering, and math. A Report from the Girl Scout Research Institute.* New York, NY: Girl Scouts of the USA (2012)
44. Munaiah, N., Kroh, S., Cabrey, C., Nagappan, M.: Curating github for engineered software projects. *Empirical Software Engineering* **22** (2017)
45. Nadri, R., Rodriguez-Perez, G., Nagappan, M.: Insights into nonmerged pull requests in github: Is there evidence of bias based on perceptible race? *Software* **38** (2021)
46. Nadri, R., Rodríguez-Pérez, G., Nagappan, M.: On the relationship between the developer's perceptible race and ethnicity and the evaluation of contributions in oss. *Transactions on Software Engineering* **48** (2021)
47. Oliveira, N., Muller, M., Andrade, N., Reinecke, K.: The exchange in stackexchange: Divergences between stack overflow and its culturally diverse participants. *Proceedings of the on Human-Computer Interaction* (2018)
48. Ortu, M., Destefanis, G., Counsell, S., Swift, S., Tonelli, R., Marchesi, M.: How diverse is your team? investigating gender and nationality diversity in github teams. *Journal of Software Engineering Research and Development* **5** (2017)
49. Parrotta, P., Pozzoli, D., Sala, D.: Ethnic diversity and firms' export behavior. *European Economic Review* **89** (2016)
50. Prana, G.A.A., Ford, D., Rastogi, A., Lo, D., Purandare, R., Nagappan, N.: Including everyone, everywhere: Understanding opportunities and challenges of geographic gender-inclusion in oss. *Transactions on Software Engineering* **48** (2021)
51. Preotiuc-Pietro, D., Ungar, L.: User-level race and ethnicity predictors from twitter text. In: *Proceedings of the 27th international conference on computational linguistics* (2018)
52. Qiu, H.S., Nolte, A., Brown, A., Serebrenik, A., Vasilescu, B.: Going farther together: The impact of social capital on sustained participation in open source. In: *Proceedings of the 41st international conference on software engineering* (2019)
53. Rastogi, A.: Do biases related to geographical location influence work-related decisions in github? In: *Proceedings of the 38th International Conference on Software Engineering Companion* (2016)
54. Rastogi, A., Nagappan, N., Gousios, G., van der Hoek, A.: Relationship between geographical location and evaluation of developer contributions in github. In: *Proceedings of the 12th International Symposium on Empirical Software Engineering and Measurement* (2018)
55. Reynolds, A., Lewis, D., et al.: Teams solve problems faster when they're more cognitively diverse. *Harvard Business Review* **30** (2017)
56. Rodríguez-Pérez, G., Nadri, R., Nagappan, M.: Perceived diversity in software engineering: a systematic literature review. *Empirical Software Engineering* **26** (2021)
57. Ross, M., Hazari, Z., Sonnert, G., Sadler, P.: The intersection of being black and being a woman: Examining the effect of social computing relationships on com-

- puter science career choice. *Transactions on Computing Education* **20** (2020)
58. Sánchez-Gordón, M., Colomo-Palacios, R.: Characterizing devops culture: a systematic literature review. In: *Proceedings of the International Conference on Software Process Improvement and Capability Determination* (2018)
 59. Sansone, L.: *Blackness without ethnicity: Constructing race in Brazil* (2003)
 60. Santamaría, L., Mihaljević, H.: Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* **4** (2018)
 61. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a "needle" in a haystack: Detecting knee points in system behavior. In: *Proceedings of the 31st international conference on distributed computing systems workshops* (2011)
 62. Sebo, P.: Performance of gender detection tools: a comparative study of name-to-gender inference services. *Journal of the Medical Library Association: JMLA* **109** (2021)
 63. Simpson, E.H.: Measurement of diversity. *nature* **163** (1949)
 64. Smith, D., Villalba, D., Irvine, M., Stanke, D., Harvey, N.: *The accelerate state of devops report* (2019)
 65. Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C., Stallings, J.: Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science* **3** (2017)
 66. Theil, H.: *Economic forecasts and policy* (1961)
 67. Trinkenreich, B., Wiese, I., Sarma, A., Gerosa, M., Steinmacher, I.: Women's participation in open source software: A survey of the literature. *Transactions on Software Engineering and Methodology* **31** (2022)
 68. Varma, R., Falk, J.H., Dierking, L.D.: Challenges and opportunities: Asian women in science, technology, engineering, and mathematics. *American Behavioral Scientist* **67** (2023)
 69. Vasilescu, B., Capiluppi, A., Serebrenik, A.: Gender, representation and online participation: A quantitative study of stackoverflow. In: *Proceedings of the International Conference on Social Informatics* (2012)
 70. Vasilescu, B., Capiluppi, A., Serebrenik, A.: Gender, representation and online participation: A quantitative study. *Interacting with Computers* **26** (2014)
 71. Vasilescu, B., Filkov, V., Serebrenik, A.: Perceptions of diversity on git hub: A user survey. In: *Proceedings of the 8th International Workshop on Cooperative and Human Aspects of Software Engineering* (2015)
 72. Vasilescu, B., Posnett, D., Ray, B., van den Brand, M.G., Serebrenik, A., Devanbu, P., Filkov, V.: Gender and tenure diversity in github teams. In: *Proceedings of the 33rd annual conference on human factors in computing systems* (2015)
 73. Vasilescu, B., Serebrenik, A., Filkov, V.: A data set for social diversity studies of GitHub teams. In: *Proceedings of the 12th Working Conference on Mining Software Repositories, Data Track, MSR* (2015). DOI <http://dx.doi.org/10.1109/MSR.2015.77>
 74. Vasilescu, B., Serebrenik, A., Filkov, V.: A data set for social diversity studies of github teams. In: *Proceedings of the 12th working conference on mining software repositories. IEEE* (2015)

75. Vasilescu, B., Yu, Y., Wang, H., Devanbu, P., Filkov, V.: Quality and productivity outcomes relating to continuous integration in github. In: Proceedings of the 10th joint meeting on foundations of software engineering (2015)
76. Wang, Y., Redmiles, D.: Implicit gender biases in professional software development: An empirical study. In: Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Society (2019)
77. Wiedemann, A., Wiesche, M.: Are you ready for devops? required skill set for devops teams (2018)
78. Wurzelová, P., Palomba, F., Bacchelli, A.: Characterizing women (not) contributing to open-source. In: 2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE) (2019)
79. Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., Skiena, S.: Nationality classification using name embeddings. In: Proceedings of the on Conference on Information and Knowledge Management (2017)
80. Zohar, I., Belmaker, M.: Size does matter: methodological comments on sieve size and species richness in fishbone assemblages. *Journal of Archaeological Science* **32** (2005)