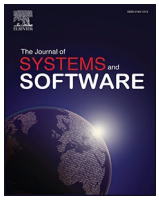
The Journal of Systems and Software 217 (2024) 112163

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss



Understanding participation and corporatization in service of diversity in free/libre and open source software development projects

Olivia B. Newton [a,*], Stephen M. Fiore [a,b]

[a] Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA
[b] Department of Philosophy, University of Central Florida, Orlando, FL, USA

ARTICLE INFO

Editor: Professor Laurence Duchien

Keywords:
Open source software
Diversity

ABSTRACT

Issues associated with a lack of diversity and inclusivity persist in the domain of free/libre and open source software (FLOSS) development and in software development generally. Researchers have suggested that the corporatization of FLOSS affords opportunities for creating an inclusive workforce. To understand the potential for firms to increase diversity, we conducted a mixed-methods study of diversity and corporate engagement in FLOSS projects. We integrate the results of a qualitative survey and a big data analysis to understand developer perceptions of corporate engagement and its association with gender and geographic diversity. In the qualitative component, we collected responses from 64 FLOSS contributors to elicit their perspectives on corporate engagement. In the quantitative component, we analyzed GitHub data from 38 projects and 9,990 contributors to investigate differences in participation and diversity based on corporate engagement. We find that contributors vary across dimensions that can inform diversity interventions: stances toward corporations and orientation towards individual/collective benefits. Our results suggest that corporate engagement may limit a project's contributor base and geographic diversity. Yet, organizations subsidizing FLOSS have opportunities to increase access to projects which would benefit diversity. This research serves to identify individual and organizational factors which may harm and help diversity initiatives.

1. Introduction

Diversity in software engineering and in free/libre and open source software (FLOSS) development in particular continues to be an area of interest as communities and organizations struggle to make significant progress towards creating inclusive and equitable workspaces. Much of the research on diversity in this work domain focuses largely on gender and/or reduces social identity to a single mode of classification (e.g., either gender or race). This is due, at least in part, to empirical evidence of a large gender imbalance in software development projects. However, recent research describes the issues associated with only considering one facet of social identity and aims to better capture the intersectional nature of experience and injustice (Winchester et al., 2022). The foundation of this work is based on the concepts of intersectionality (Crenshaw, 2018) and multiple jeopardy (King, 1988) which describe the multifaceted nature of social identity to account for the interdependence of systems used to dominate and discriminate against marginalized groups. It is therefore important for research in this topic area to examine multiple aspects of social identity, to better capture the variety of experience and potential diversity in FLOSS development.

In calls to improve inclusivity, researchers have suggested that the increased engagement of corporations in FLOSS projects may potentially be leveraged for greater good. These researchers assert that the transformation of this form of software development has far-reaching effects as significant changes are made to the nature of work (Germonprez et al., 2019). Furthermore, they prompt researchers to consider shifts in FLOSS governance and the potential effects of the transformation on diversity, inclusivity, and equity. These are topical issues for FLOSS development with implications for participation and sustainability in technological production.

According to the situated knowledge thesis put forth by Haraway (1988), "what one knows or experiences reflects one's social, cultural, and historical location" (Schmaus, 2015, p. 245). As such, the inclusion of groups that have been excluded from software development has implications for broader societal issues and labor questions. Inclusivity is related to equity in knowledge production (whose knowledge is valued when building technologies), accessibility and usability of technology (who can use the technologies), and the direction and application of

* Corresponding author.
E-mail address: olivia.newton@ucf.edu (O.B. Newton).

https://doi.org/10.1016/j.jss.2024.112163
Received 8 November 2023; Received in revised form 26 June 2024; Accepted 20 July 2024
Available online 26 July 2024

technological advances (which problems technologies are designed to address). Research has demonstrated that individuals who differ along social dimensions, like gender, also differ along other dimensions that influence problem-solving processes and outcomes in software development (e.g., information processing style, Gralha et al., 2019); knowledge sharing patterns, Zolduoarrati et al., 2022). The effects of exclusivity in software development are evident in the many cases that software has failed some populations (Hilderbrand et al., 2020).

Understanding the need for greater inclusivity and equity in technology development, and recognizing the potential of organizations, we designed a mixed-methods study to model the relationship between corporate engagement and social diversity in FLOSS projects. The goal of this work is to (1) understand how FLOSS contributors respond to corporate involvement and (2) analyze the association between corporations and diversity in project ecosystems. Our research includes analyses of qualitative survey data and GitHub project data which we integrate towards this goal. The results of this research provide insights regarding participation decisions and levels of diversity observed in FLOSS projects.

First, we find that FLOSS contributors can be differentiated on the basis of their stance towards corporations, which influences their decisions for initial and sustained participation in projects. Second, we find that FLOSS contributors largely view the effects of corporate engagement on diversity as either neutral or positive. Third, we observe significant differences in likelihood to sustain participation in projects with and without corporate engagement, with contributors to the latter exhibiting greater commitment early in their participation. Lastly, we observe that projects with corporate engagement generally have the same levels of gender diversity as projects without corporate engagement. We also observe that projects without corporate engagement tend to have higher levels of geographic diversity, demonstrating more openness across national boundaries.

We next summarize research on corporate engagement and participation in FLOSS to situate the present study. We follow this review of related work with a detailed description of our approach, results, and contribution to this research area. From this, we consider opportunities for studying and increasing inclusivity in service of more equitable participation in FLOSS development. The research presented here builds on understanding of the role and capacity of organizations, public and private, for diversity and inclusivity in the space of FLOSS development.

2. Related work

2.1. Corporate engagement in FLOSS development

Corporations initially rejected the free and open source paradigm, but, since the mid-2000s, they have maintained a relationship with FLOSS communities, altering development processes and licensing along the way (Fitzgerald, 2006). Research has thus been conducted across disciplines to model the relationship between corporations and communities in an effort to build foundational knowledge (descriptive works) and, in some cases, to provide recommendations for the prosperity of such relationships (prescriptive works). Different types of corporation-community relationships have been studied in FLOSS development. Findings suggest that a corporation's management style can introduce friction with communities, and that corporations and communities hold different types of power.

In early research, Dahlander and Magnusson (2005) proposed three types of firm-community relationships that fall along a continuum of community benefits: *symbiotic* (firm gains-community gains), *commensalistic* (firm gains-community indifferent), and *parasitic* (firm gains-community loses). The symbiotic approach, in which both the firm and the community benefits from the relationship, is associated with a unique set of managerial challenges but also produces the greatest influence over the community. In this approach, the firm must carefully consider the nuances of the community: organizational decisions should not conflict with community norms, values, and goals. The other approaches—commensalistic and parasitic—are understood to afford little to no influence over the community. In these approaches, the corporation-community relationship is untenable, as the firm's use of communal resources without reciprocity results in a negative reputation and lost ties (Dahlander and Magnusson, 2005).

With increased corporate engagement, corporate structures and control have become prevalent in the space of FLOSS development. Scholars have identified this corporatization of FLOSS as diminishing the public good associated with the paradigm. This is due to the fact that corporate goals fundamentally conflict with FLOSS's orientation towards public goods and undermine its sociopolitical power (Birkinbine, 2015). The idealized FLOSS development project favors equality, sharing, and reciprocity with the aim of fostering innovation and advancing shared resources (Germonprez et al., 2013; Rajanen and Iivari, 2015). In contrast, corporatization favors efficiency and commercial success. Published works on these topics delineate the harms imposed on projects by firms. For example, firms impose control structures (e.g., licensing) that may inhibit software freedom and openness, and changes in governance (e.g., through acquisitions) can derail project plans previously developed in collaboration with a community (Birkinbine, 2015). At the same time, there are methods that can be used by communities to maintain power and protect the products they contribute to the commons (e.g., by forking projects; Birkinbine, 2015). In sum, these studies highlight the ways that firms alter the goals of FLOSS development and how communities may respond to them.

In the literature, research continues to examine the relationship between communities and the organizations that collaborate with them (Germonprez et al., 2019; Germonprez et al., 2017). The levels at which corporations engage with FLOSS development have been elaborated to include: economic and business markets for the software; communal resources for projects and contributors; and "social and material rules" in FLOSS communities (Germonprez et al., 2013). Collaborations between organizations and communities have resulted in a shift away from the development of standard platforms to a focus on usability and features (Germonprez et al., 2017). Despite the increasing attention paid to corporatization of FLOSS, it is unclear if and to what extent firms have constrained or facilitated positive social outcomes (e.g., more equitable technologies or inclusive communities; Newton, 2020). Furthermore, motivating FLOSS contributors to accept organizational initiatives for inclusivity and equity necessitates an understanding of their attitudes towards corporate engagement in this space. This is significant given that FLOSS contributors differ in terms of sociopolitical ideology, from anarchism to nationalism, to anti-authoritarianism (Coleman, 2013). We next turn to research on the composition and behavior of FLOSS contributor groups.

2.2. Participation and diversity in FLOSS development

Much research has been conducted on participation in FLOSS development. Research on this topic has been made increasingly feasible through the extraction and analysis of contribution behavior from online platforms like GitHub and corresponding version control systems. Contributors to FLOSS projects have been characterized and differentiated from each other in a number of ways. This has included the classification of FLOSS contributors on the basis of their activity levels (e.g., high versus low), status and influence (e.g., technical role and/or popularity), and project commitment (e.g., paid versus volunteer). Demographic information, including gender and geolocation, has also been used to identify differences in participation between project contributors, showing that significant disparities exist with respect to *who* participates and *how* they participate. Notably, the population of FLOSS contributors has low social diversity across a number of dimensions like gender, race, and geographic location (El Asri and Kerzazi, 2019; Diversity 2022).

Research on diversity in FLOSS projects has largely focused on

gender. Studies have been conducted on (1) the presence, or lack, of women; (2) the tenure of women; (3) the biases reported by women; (4) the bias against women observed in projects; and (5) the experiences and behaviors of women in FLOSS development spaces. Generally, it is observed that women participate in FLOSS projects at lower rates than men (David and Shapiro, 2008; Vasilescu et al., 2015) and have shorter tenure in projects and online platforms (Qiu et al., 2019; Newton and Song, 2022). Some research has explored the presence and effects of bias experienced in FLOSS projects. This work is insightful given that FLOSS contributors have some level of awareness of potential collaborators' gender in online platforms (Vasilescu et al., 2015) and the salience of gender can influence interactions in social spaces (McNicol, 2013).

The rate at which women's contributions are accepted varies on the basis of both the salience of their gender and their membership status in FLOSS projects (Terrell et al., 2017). When a woman's gender is salient, and she is not a project "insider," her contribution is less likely to be accepted in comparison to a male peer whose gender is salient and is also not an insider (Terrell et al., 2017). This relationship between gender and contribution integration is flipped when gender is obscured: a woman is more likely to have her work accepted when compared to a male peer. Such a finding suggests that the difference in contribution evaluation may be the result of bias rather than level of competence. Other research on gender shows that women tend to contribute to a smaller set of projects compared to men and are more reserved in communication, specifically engaging in fewer emotional displays and using less profanity (Imtiaz et al., 2019). Women themselves have reported experiencing discomfort and harassment in FLOSS development spaces (Nafus, 2012). Perhaps unsurprisingly, it is well established that gender diversity is relatively uncommon in FLOSS development (e.g., Vasilescu et al., 2015). These works show that the interplay of social (e.g., norms and bias) and technical factors (e.g., tenure and experience) in FLOSS projects are associated with gender imbalances in participation.

A recent study suggests that contributors whose race is perceptible and underrepresented tend to have less work accepted in FLOSS projects (Nadri et al., 2021). At least one survey of developers revealed that conflicts emerge as a result of differences in language, nationality, and ideology between collaborators (Vasilescu et al., 2015). Less work has accounted for variation in other dimensions (e.g., socioeconomic status, education, and age) or intersections between them. Turning to group and organizational factors, there is some evidence that FLOSS volunteers' decision to participate in and commitment to a project is influenced by social norms and the project's association with positive causes (Barcomb et al., 2018). At the group level, different types of diversity, namely tenure and gender, have been empirically linked to turnover in FLOSS projects (Vasilescu et al., 2015). But, to date, the relationship between organizations, participation decisions, and diversity in FLOSS projects has received less attention from researchers. We see this gap as an opportunity to identify potential avenues for improving inclusivity and equity in FLOSS development.

Germonprez et al. (2019) speculate that social computing technologies may hold a space of particular importance in addressing problems emerging in this work domain. Social computing technologies are a type of collaboration infrastructure that is created and designed by organizations, including, for example, social coding platforms like GitHub. Given the embedded nature of these types of technologies in software development, they are a critical area for understanding individual, group, and organizational behavior relevant to diversity and inclusivity. Based on this, we identified GitHub as a highly relevant and appropriate space for our research on these phenomena. We now turn to our study of corporate engagement and diversity in FLOSS projects on GitHub.

3. Study approach

In this section, we describe our research goals and approach. Our research was motivated by an interest in clarifying the relationship between corporate engagement, diversity, and participation in FLOSS projects. More specifically, our goal is to characterize FLOSS contributors' perceptions of corporations as they relate to participation and diversity to better understand if and how the presence of these organizations may be leveraged to improve diversity and inclusivity. We argue that corporations can and do influence participation and group formation in FLOSS development. Their influence on these phenomena is achieved through a variety of mechanisms. Firms can directly alter participation through recruitment and assignment of developers to FLOSS projects. Given the variety of values and goals held by FLOSS contributors, firms may also alter participation by their mere presence, attracting contributors based on affinity or repelling contributors based on sociopolitical beliefs.

We therefore solicited information about corporate engagement in FLOSS development from project contributors, including the relationship between corporations and diversity. With the data collected, we employed a cross-sectional approach to examine differences in responses based on sociodemographic information. We complemented our survey research with an analysis of diversity, participation, and corporate involvement in FLOSS projects hosted on GitHub. This mixed-methods approach enabled us to provide a more complete picture of study concepts by capturing both individual and group factors, and the differences and similarities between different types of projects. The qualitative component allowed us to build foundational understanding of contemporary individual attitudes and beliefs on corporate engagement in FLOSS development. The quantitative component allowed us to examine group behavior and composition in FLOSS projects based on corporate engagement. Together, they illustrate how corporate engagement influences individual behavior to produce group differences. This study was thus completed in two parts–survey research and GitHub project data research–and we organize the paper accordingly.

3.1. Research questions

The research presented here was guided by the following questions:

RQ1 Perceptions of Corporations: What are contributors' perceptions of corporate involvement in FLOSS projects?

(a) *Experience*: Do contributor experiences in projects differ based on corporate involvement?
(b) *Participation*: Does corporate involvement influence contributor intent to sustain participation in a FLOSS project?
(c) *Diversity*: Do contributors believe that corporate involvement helps or hurts the achievement of diversity in FLOSS development?

RQ2 Effects of Corporations: Do quantitative differences exist between projects with and without corporate involvement?

(a) *Participation*: Is corporate involvement associated with differences in sustained participation in FLOSS development?
(b) *Diversity*: Is corporate involvement associated with differences in social diversity in FLOSS development?
    i. *Gender*: Are differences in gender diversity associated with corporate involvement?
    ii. *Geographic*: Are differences in country diversity associated with corporate involvement?

The combination of the above questions provides us with a better understanding of the individual (participation decisions) and organizational (corporate engagement) factors which produce differences in group phenomenon (social diversity in FLOSS development). We next provide the details of the mixed-methods research conducted to answer these questions.

4. Methods

Data were collected and analyzed from (1) contributors to FLOSS development on GitHub and (2) FLOSS projects on GitHub to elucidate the relationship between corporate involvement, participation, and diversity. Supplemental materials are provided via the Open Science Foundation[1].

4.1. Data collection: survey research

We administered a survey via Qualtrics to collect data on contributor perceptions of corporate involvement in FLOSS projects. The inclusion criterion was contribution to FLOSS development on GitHub. Participants were provided with an explanation of research approved by our institution's Institutional Review Board and asked to provide confirmation of their decision to voluntarily participate in the study. Participants who completed the survey were compensated for their time and effort with a $5 Amazon.com gift card. On average, the survey took approximately 15 minutes to complete. The survey was open to responses from May 2022 to September 2022.

The survey was administered in two phases: a general call for participation distributed via social media and listservs (phase 1 - broad distribution) followed by a targeted invitation to participate based on contribution data extracted from FLOSS projects on GitHub (phase 2 - targeted distribution). In the first phase, the survey was distributed via Twitter, the GitHub Subreddit, and a set of listservs: the NIH Science of Team Science listserv, the Listserv For Scientists, the Free Software Foundation's LibrePlanet listserv, and the International Federation for Information Processing's Working Group on Open Source Software listserv. This distribution resulted in over 400 responses, but as is common with surveys distributed via social media (Xu et al., 2022), many responses were unusable due to evidence of automated activity and/or incomprehensible, incoherent, or plagiarized text responses. We thus kept 34 complete responses to the survey in the broad distribution.

In the second phase, an invitation to participate was sent directly to the contributors of FLOSS projects selected for the GitHub project research component of this study. A total of 596 contributor emails were extracted from commits and pull requests. Of these emails, 10 were returned as undeliverable. This distribution resulted in 57 responses (10% response rate); However, 22 data points were labeled incomplete and unusable due to the fact that the participants responded to less than 10% of the survey and 0% of the open-ended questions. 5 additional participants reviewed the explanation of research but did not complete the survey. As a result, 30 complete responses to the survey in the targeted distribution. Our final sample consisted of 64 participants across the two survey distributions. Demographic information for the survey sample is provided in Tables 1, 2, 3, 4.

4.2. Data collection: GitHub project research

Data were collected and aggregated from the GHTorrent mysql-2019-0601 data dump for sample selection which included data for over 125 million repos. Additional data were extracted from the GitHub REST API using the PyGitHub library[2] and GrimoireLab's Perceval module[3]. Only repositories that were devoted to cooperative software development projects and were active at the time of data collection were classified as belonging to the study population. Specifically, GitHub repos were included in the study population if they met the following criteria:

1. *Software development focus*: a specified programming language;
2. Continuity *of development activity*: at least 5 years old;
3. *Sufficient amount of development activity*: at least 50 commits;
4. *Sufficient number of contributors*: at least 5 members.

Based on this, the study population was made up of 2572 repositories. This is approximately 0.002% of all the repos in the GHTorrent data dump.

To select a sample of projects for analysis, project data were aggregated across a set of dimensions. This set of dimensions was selected on the basis of their relevance to the research topic (Nagappan et al., 2013) and prior research on project growth and attraction of newcomers (Fronchetti et al., 2019). For empirical research on participation in FLOSS projects on GitHub, the space consists of the following dimensions: number of contributors (12-month period), number of commits (12-month period), number of forks, number of watchers, main programming language, project age, and committer turnover (comparison of two consecutive 12-month periods). These aggregated data then served as input for a sample coverage algorithm (Nagappan et al., 2013) to select 20 GitHub projects. Two projects were removed from the sample: one is a mirror of a repo hosted outside of GitHub, and the other had been deleted by the time of data collection (September 2022). The remaining 18 projects were kept for analysis. Summary statistics used to compare the study population and sample along the aforementioned dimensions is provided in online supplemental materials.

Table 1
The gender makeup of the survey research sample.

Response	N_{social}	$N_{committers}$	Overall (%)
Man	30	24	54 (84%)
Woman	3	1	4 (6%)
Prefer not to say	1	4	5 (8%)
Nonbinary	0	1	1 (2%)

Table 2
Descriptive statistics for age reported by survey participants.

Age	N_{social}	$N_{committers}$
Minimum	21	20
Maximum	44	52
Mean (±SD)	29.51 (±5.81)	34.13 (±8.95)
Median	28	31.50

Table 3
Geographic location reported by survey participants.

Continent	N_{social}	$N_{committers}$	Overall (%)
Americas	31	12	43 (67%)
Asia	1	4	5 (8%)
Europe	2	13	15 (23%)
Oceania	0	1	1 (2%)

Table 4
The racial and ethnic makeup of the survey sample.

Race/Ethnicity	N_{social}	$N_{committers}$	Overall (%)
Arab, Non-White	0	1	1 (2%)
Asian	5	3	8 (12%)
Black	4	0	4 (6%)
Hispanic/Latine, Non-White	0	2	2 (3%)
North African	0	1	1 (2%)
White, Arab	0	1	1 (2%)
White, Hispanic/Latine	1	2	3 (4%)
White, Non-Hispanic/Latine	24	19	43 (67%)
White, Pacific Islander	0	1	1 (2%)

[1] OSF project link

[2] https://pypi.org/project/PyGithub/

[3] https://github.com/chaoss/grimoirelab-perceval

We then expanded this data set to additionally consider other projects that constituted the ecosystems of the 18 projects. We applied the reference coupling method proposed by Blincoe et al. (2015) to identify projects belonging to ecosystems of the 18 sampled projects. This method is based on mentions of dependencies between projects in comment data extracted from platform API. We scraped comments linked to commits, pull requests, and issues. The cross-reference patterns reported in Blincoe et al. (2015) were used: User/Project#Num (e.g. rails/rails#123) or User/Project@SHA. Extracted patterns were manually inspected to remove false positives and self-referential matches. Of the 18 projects in the sampled data, cross-references to other projects in comments were found for nine projects. These nine projects and the projects in their ecosystem were used to analyze participation and diversity differences. Our final data set included information for 38 projects, 9990 contributors, and 445,465 events (commits, pull requests, and issues).

These data were prepared based on best practices for studying contributor behavior in FLOSS projects. As in prior research on GitHub (e.g., Golzadeh et al., 2021), bots were identified and removed from the data set based on username, information in GitHub profile bio, and/or GitHub's contributor type label (user or bot). All usernames and bios that contained the patterns 'bot', 'CI', 'automated'/'automation', or 'machine' were extracted in an iterative process (e.g., all accounts selected in 'bot' pattern matching were removed for closer inspection before beginning 'CI' pattern matching). Then, extracted accounts were manually inspected to remove false positives (e.g., 'bot' pattern in a human's username). Through this process, a total of 30 accounts were labeled automated and removed from the data set. Some of these accounts were found across multiple projects (e.g., dependabot) whereas others were specific to a project.

Usernames were used to extract location information provided in GitHub profiles. Location information was prepared for analysis using the tmaptools (Tennekes, 2018) and tidygeocoder (Cambon et al., 2021) packages in R. Functions in these packages were used to define coordinates for 6834 contributors with location information in the user profile. The coordinates were then used to assign country and continent-level labels to each contributor. 98 countries were represented in the data with the top ten being United States, Germany, China, United Kingdom, India, France, Canada, Netherlands, Poland, and Switzerland. 8 countries were represented by a single contributor: Brunei, Dominican Republic, Nepal, Panama, Paraguay, Saudi Arabia, Sudan, and Venezuela.

Identity merging was applied using the script[4] by Vasilescu et al. (2015). The approach first merges user aliases if they share the same email address and then evaluates name, username, email address prefix, email address domain, and location information. In line with the authors' guidance on the application of their approach, we manually inspected outputted files. To limit the number of false positives (i.e., aliases incorrectly merged; El Asri and Kerzazi, 2019), we have been as conservative as possible when deciding to merge aliases. The genderComputer tool[5] was used to infer the gender of project contributors based on name and location data. In previous research reporting the development of genderComputer tool, this gender resolution technique was found to have precision of up to 93% (Vasilescu et al., 2015).

4.2.1. Accuracy of gender resolution

The genderComputer tool resolved gender for 8349 (84%) of users in the sample: 7608 (76%) contributors were labeled as man and 741 (7%) as woman with the remaining 1641 (16%) unlabeled. The output of the tool was manually inspected for a random selection of 500 labeled contributors, and compared against information in GitHub and social media profiles. Less than 15 instances of classification errors (0.001%) were discovered in this process. Such errors were produced as a result of nicknames (e.g., Oli, Dani, and Juaky) and gender-neutral names (e.g., Taylor and Casey).

We also evaluated the match between the genderComputer labels against pronouns specified by contributors in GitHub profiles. Pronouns for contributors were collected using a GitHub GraphQL query. 698 contributors (7% of the sample) included information in the pronouns field of the profile, with 691 contributors using it to share pronouns. The other 7 contributors had non-pronouns in this field (e.g., "Attack Helicopter" and "Yes/No"). We used contributors' specified pronouns to label gender. 613 contributors with only he/him/his pronouns were labeled as man, 46 contributors with only she/her/hers pronouns as woman, and 32 contributors with a blend of gendered and/or gender-neutral pronouns were unlabeled. We next compared pronoun-based labels with genderComputer labels. 87 contributors (79 man, 3 woman) who had specified pronouns did not have a genderComputer label as the tool was unable to resolve their gender. This set was not included for evaluation of the tool. Data for the 604 contributors were used to calculate evaluation metrics for the genderComputer tool, with gender labels based on specified pronouns serving as ground truth (Table 5). We observe that the tool performs well, and more so for men than women. A total of 73 (12%) contributors were misclassified (i.e., had a mismatch between pronoun-based label and tool-based label). 26 of these mismatches were cases in which a contributor specified a blend of gendered and/or gender-neutral pronouns.

4.3. Data analysis: survey research

To analyze participant responses to Likert-type items, aggregation and visualization techniques were applied to the data using R and in particular the tidyverse (Wickham et al., 2019) and likert (Bryer and Speerschneider, 2022) packages. These data summaries enabled us to examine the distribution of experiences as represented in the survey data. Reflexive thematic analysis (RTA), an approach for thematic analysis elaborated on by Braun and Clarke (2019), was applied to participant responses to open-ended survey questions to characterize FLOSS contributor experiences and dispositions regarding study concepts. This approach is flexible and acknowledges the researcher's "active role in knowledge production" [Bryer, 2022, p. 1393]. The six recursive phases of RTA are: (1) familiarization with the data; (2) generation of initial codes; (3) generation of themes; (4) review of potential themes; (5) definition and naming of themes; and (6) report production. Details about the RTA applied to participant responses to the survey are provided in Table 6.

We next address underlying theoretical assumptions associated with our application of RTA. We adopted a constructionist epistemology: our understanding of reality, experiences, and language is socially produced and reproduced through interaction within and outside ourselves (Burr, 2015). Understanding is "historically and culturally relative" [Burr, 2015, p. 4]. Thus, although we attended to the recurrence of codes, it was not the most significant criteria applied to the data when constructing themes. Instead, we primarily weighed meaning and meaningfulness as the "central criteria" in coding [Byrne, 2022, p. 1395]. Our analysis had a primarily experiential orientation, reflecting the feelings and experiences of participants (Clarke and Braun, 2014). In considering implications and potential future work based on responses, a shift towards a critical orientation was necessary to interpret them in relation to the social context in which our participants are embedded (Clarke and Braun, 2014). That is to say, each participant has a sociocultural position

Table 5
Classification evaluation metrics for genderComputer tool.

Precision	Recall	Accuracy	F1
0.97 [0.95, 0.98]	0.94 [0.92, 0.96]	0.92 [0.89, 0.94]	0.96

[4] https://github.com/bvasiles/ghtunmaskingaliases
[5] https://github.com/tue-mdse/genderComputer

Table 6
Summary of Reflexive Thematic Analysis Phases. Codes and themes iteratively produced and refined in the RTA to represent differences experienced by participants following their response to, "Does your contribution experience differ for FLOSS projects with corporate involvement and FLOSS projects without corporate involvement?" and "How much does the lack or presence of corporate involvement influence your decision to continue participating in a project after initial contribution?".

RTA Phase Description and Findings
Data Familiarization: initial line-by-line review of responses. Participant responses described specific and general differences they have experienced and observed over time in FLOSS projects.
Initial Coding: 89 codes produced in second line-by-line review. Open coding applied to participant responses with the aim of capturing experiences and perspectives. Most frequent codes include: motivation, time, bureaucracy, and organization.
Generating and Reviewing Themes: 17 themes produced. Codes reviewed to identify relationships between them and form themes for projects with and without corporate involvement. Codes describe the effect of corporations on contributors and projects as a whole.
Defining and Naming Themes: grouped themes into 4 major areas. Themes and sub-themes reflect stances towards corporations in FLOSS development (positive, negative, neutral, and mixed) and orientation with respect to potential individual and collective benefits.

which, like ours, interplays with understanding and its production. Lastly, we conducted an inductive (i.e., data-driven) analysis in which codes were descriptive and based on explicit meaning (Byrne, 2022; Clarke and Braun, 2013).

4.4. Data analysis: GitHub projects research

As a first step, we supplemented the data with information from READMEs, project websites, and Wiki pages. This was done to categorize projects into one of three groups: no corporate influence, receives corporate support, owned by a corporation (Table 7). We refer to these groupings as project type. This qualitative categorization of projects was used to analyze group differences. All subsequent computations and analyses were conducted in R. The corresponding data and code files are provided in the online supplemental materials

4.4.1. Sustained participation

To compare sustained participation between project types, survival analysis was applied to contributor data. This analysis serves to complement findings from the survey research on experiential differences and participation decisions. Specifically, it allows us to observe the group-level patterns of participation that emerge in association with the interaction between project structures, project processes, and individual motivations. Survival analysis used is across disciplines for the evaluation of event rates and causal mechanisms (Liu, 2012). In research on FLOSS development, survival analysis has been used to model software releases (Sen et al., 2015), assess development processes (support quality; Jarczyk et al., 2018) and quantify gender differences in participation (Qiu et al., 2019).

Survival analysis allows for the assessment of time until an event occurs, which in this case is a contributor's disengagement from a project. Importantly, for some contributors, we do not know if and when they will disengage from a project in the future. We therefore have *censored* data, or partial observations for a subset of contributors (Gijbels, 2010). In our case, the data is right censored (Gijbels, 2010; Clark et al., 2003) by the event of data collection (i.e., *when* we scraped the data): a contributor may have left a project at a date which occurs after data was collected for this research. Survival analysis handles censored data directly, making it appropriate for use with our data (Clark et al., 2003). We calculate Kaplan-Meier survival estimates (Clark et al., 2003; Kaplan and Meier, 1958) of the probability of remaining in a project over time. We use survival probabilities to evaluate differences in sustained participation between projects with and without corporate involvement. We implemented the survival analysis and applied significance testing in R using the survival package (Therneau, 2024), and plot survival curves using the ggfortify package (Tang et al., 2016; Horikoshi and Tang, 2018).

Table 7
Levels of corporate involvement in FLOSS project sample.

Level	No. Projects
No clear corporate influence	12
Receives corporate support	11
Owned by a company	15

4.4.2. Gender and geographic diversity

To evaluate differences in social diversity as they relate to project type, the Blau index was computed for gender and country in each project. This analysis complements findings from the survey research on perceptions of corporations' effect on diversity in FLOSS. Specifically, it allows us to examine if FLOSS contributor perceptions align with observed levels of diversity in projects with and without corporate involvement. The Blau index was developed by Simpson (1949). This operationalization of diversity measures the variety of a categorical data type within a unit (Harrison and Klein, 2007). The formula for the Blau index is $1 - \Sigma p_k^2$, "where p is the proportion of unit members in kth category" [55, p. 1211]. The values computed range from 0 and $(K - 1)/K$.

For our calculation of gender diversity, the maximum possible value is 0.50 or 0.67, depending on the data used as input. The first Blau index value for gender diversity has a maximum possible value of 0.50 as it is based on the two gender groups we were able to label using genderComputer. A value of 0.50 for gender diversity represents complete evenness in the presence of the two gender groups within a project. The second Blau index value for gender diversity has a maximum possible value of 0.67 as it is based on the genderComputer labels enhanced with pronoun information. Specifically, it accounts for pronoun-based corrections to the two gender groups labeled by the tool in addition to a third gender group for contributors with gender-neutral and/or blended pronouns who were previously unlabeled. For this value, 0.67 represents complete evenness in the presence of these three gender groups within a project. For our calculation of geographic diversity, the maximum possible value is approximately 1.00 as there are 98 countries represented in the data (out of the 195 in the world according to the United Nations;Saunders, 2024). The richness of the data for geographic diversity allows for a higher maximum value compared to gender diversity in our analysis. To determine if there was a statistically significant differences between project types, models were created with diversity as the response and corporate involvement as predictor. A non-parametric approach—Welch's unequal variances t-test—was applied due to the skewed distributions observed in the data. We report descriptive and test statistics from this analysis, and visualize distributions using density plots.

5. Results

To answer our first set of research questions (RQ1a-c), we characterize contributor perceptions of corporations based on their responses to Likert-type items and open-ended questions. We specifically cover perceptions of contribution experience and the effect of corporate engagement on diversity in FLOSS projects (Figs. 2–5). To answer our

second set of research questions (RQ2a-b), we analyze differences in participation (Fig. 6, Table 9 and diversity (Figs. 7, 8) in FLOSS projects based on GitHub contribution data and project type (level of corporate involvement).

5.1. Survey research: experiential differences

With respect to **RQ1a**, the results of the survey suggest that there are considerable differences in contribution experience between projects with and without corporate involvement. These differences vary somewhat between genders across geographical boundaries (Figs. 1, 2). For example, the majority of participants who were located in the Americas and Asia tended to report experiencing at least some difference between projects with and without corporate involvement (Fig. 2). Furthermore, over half of participants in the Americas reported experiencing differences in contribution most of the time or always. In contrast, few participants in Europe reported experiencing differences most of the time or always in their contributions to FLOSS projects.

5.1.1. RTA results: elaborations on experiential differences

When asked to explain the differences experienced between projects with and without corporate involvement, most survey participants framed their experiences in *positive* and/or *negative* terms, and a small number reported experiencing no difference. Four major themes regarding differences between project types were produced in the RTA:

1. *Project structure as restriction and independence*,
2. *Project process as resource*,
3. *Problems with corporations as incongruence*, and
4. *Motivations of contributors as congruence*.

Within the themes of structure and motivation, participants identified unique attributes of projects with corporate engagement and projects without corporate engagement. The themes and subthemes of process and problems were specifically focused on projects with corporate involvement.

In the *project structure* theme, participants expressed dissatisfaction with both corporate and non-corporate projects (i.e., negative aspects), and praised the natural fit of non-corporate structure with the values of FLOSS (positive aspects). The structure of projects with corporate involvement was viewed as controlling and cumbersome, resulting in an inflexibility around completing task work due to top-down constraints. Further, there is a need for increased effort on the part of contributors to determine the appropriate sequence of actions and meet bureaucratic demands: *"they'll often require CLAs [Contributor License Agreements] and other burdensome impediments"* [GH14]. Participants stated that this extended the time to complete tasks by imposing unnecessary requirements and were likely the result of the software not being a priority for the associated company. As one participant noted:

> "*most of the time, projects with corporate involvement tend to lead to a lower quality of work and more time consumed as corporations set up initiatives and guidelines we all have to follow which slows everything down*" [SM2].

Although fewer participants focused on FLOSS projects without corporate involvement when elaborating on differences in structure, those who did described the lack of structure as producing both benefits and consequences. On the positive side, the governance models of non-corporate projects are perceived as fairer and granting greater freedom to contributors. This allows them to improvise when producing code that addresses their own and others' needs. On the negative side, the lack of structure can create coordination issues: *"projects without corporate involvement are more public, but also more disorganized, with mostly no one doing project management"* [GH4]. The structure of corporate projects is perceived as highly restrictive, whereas the structure of non-corporate projects is perceived as affording independence with the consequence of disorganization.

In the *process* theme, participants described positive experiences in projects with corporate involvement. A common subject across participants was the speed and quality of support in associated projects: *"projects with corporate involvement are usually lively and faster"* [SM9] and *"it feels more professional"* [GH24]. Projects with corporate involvement also potentially make better use of features provided by the online platform (GitHub) to coordinate work and further have resources to share information relevant to coordination requirements:

> "*Corporate-backed FLOSS projects usually have a more organized bug tracker. New issues are triaged quickly. Current development priorities ("sprints") are sometimes public*" [GH27].

Other participants similarly noted that documentation for processes supported their ability to contribute and work with project integrators. Corporate engagement in FLOSS projects thus introduces a trade-off for contributors: while corporations may be associated with faster response times and greater information transparency from managers and integrators, they are also associated with increased time and effort on the part of code contributors with no *perceived* benefit to software quality.

In the *problems* theme, factors which produced distinct experiences include potential risks of joining projects with corporate involvement and the goals of companies involved in FLOSS. A set of risks identified by survey participants exists at the project level: abandonment, corporate wrongdoing, scale, and conflict. These first two risks were seen as potentially producing the same effect on project success:

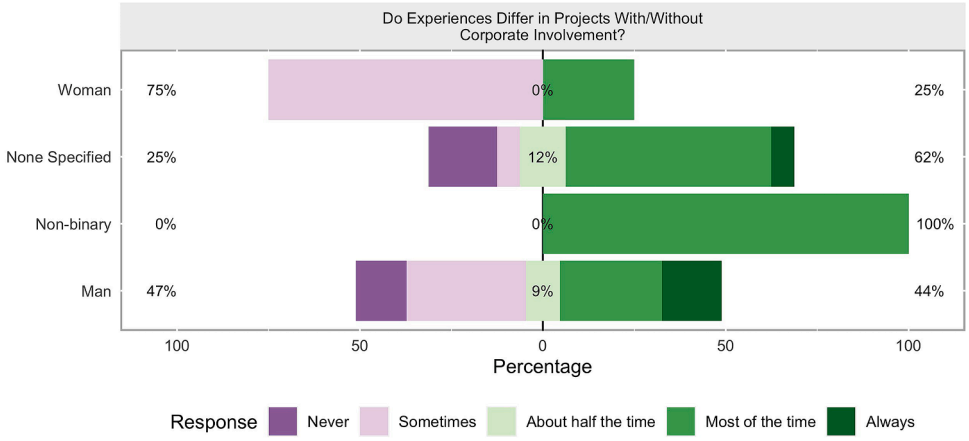


Fig. 1. Experiential difference in projects with and without corporate involvement, broken out by participant reported gender.

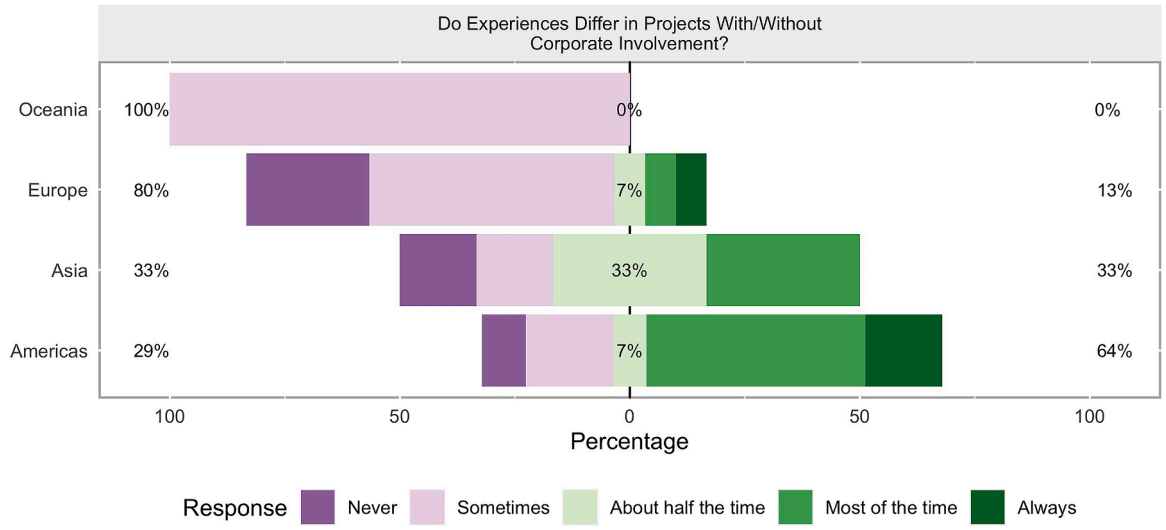


Fig. 2. Experiential difference in projects with and without corporate involvement, broken out by participant reported location (at continent level).

"*Because a corporation may be involved in a FLOSS project, there is a concern that if the corporation in question is involved in nefarious dealings (or voluntarily abandons support) that it can severely damage or limit the momentum of a project*" [GH6].

The third type of risk similarly reflects a concern about insufficient resources, but in the opposite case: a FLOSS project with corporate involvement may become popular and grow significantly without the personnel or infrastructure to meet needs at scale. Finally, the fourth risk is linked to control and the foundations of FLOSS, namely licensing, in which participants expressed concerns about legal issues that may emerge from contribution to projects with corporate involvement. The goals of firms involved in FLOSS were also seen as producing problems, due to execution of decisions without community input or explicitly disregarding community input. In such cases, there exists a misalignment between community and corporate goals as a firm's actions reflect their own priorities and power asymmetries: *"there is sometimes priority given to 'get it done' or 'this is the way we've decided' rather than understanding community feedback"* [SM34] and *"the corporate contributors usually have their own agenda"*[GH22].

In the *motivations* theme, responses were entirely positive regardless of their focus on corporate or non-corporate projects. In describing how differences between project types, some participants shared that they gained individual experiential value when contributing to projects with corporate involvement, whereas others shared that they either gained the opportunity to contribute to advancing software and programming more broadly or to engage in activity that brought them joy with like-minded others. Here, the distinction between project types was very clear.

Positive aspects of projects with corporate involvement reflected personal benefits—career opportunities and compensation—in addition to satisfactory work support processes and characteristics. Such projects provide early career professionals with the opportunity to build relationships for networking purposes, gain work familiarity with developers and tasks, and participate in activities outside the project: *"gave me great experience with the devs and managers"* [SM1] and *"creates opportunities to participate in open source internships or retreats from time to time"* [SM23]. Another important component of corporate involvement for FLOSS contributors is the potential for compensation. Positive aspects of projects without corporate involvement in contrast reflect collective benefits—better code and products—in addition to outlets for joy and the expression of social values. Contribution in these projects may be more *"like a hobby"* [GH20] which is related to *"more freedom [and] more intrinsic motivation"* [GH8].

5.2. Survey research: corporations and participation decisions

With respect to **RQ1b**, responses covered a number of reasons for sustaining or avoiding participation in corporate-involved projects. In general, participant responses fit within the themes generated in the analysis of experiential differences between types of projects (Fig. 3), with the majority belonging to *problems with corporations as incongruence* and *motivations as congruence*. Under the *problems* theme, concerns centered on the possibility that contributions would become restricted as the intellectual property of the company in question. Among reasons to avoid or leave a project, participants expressed that they experienced discord with decision makers and a misalignment of values. Concerns about project-specific goal misalignment were expressed by survey participants: *"presence of the corporate involvement makes me not want to get further involved since the main leads try to alter the whole project to fit their (corporate's) vision for the project"* [SM12].

Under the *motivations* theme, some contributors are driven to sustain participation due to perceived personal benefits (e.g., compensation, network, etc.), while others are demotivated by a lack of collective need and social good:

"*I'm less interested in helping with projects that already have corporate funding since they have the money to hire someone, while other projects I could work on don't*" [SM15].

"*It's more of a moral dilemma - Like why should I contribute to the benefit of a corporate whose only motivation is to become richer? I would rather spend my time with FOSS projects whose contributors are driven by passion for programming (most of them) and the prime motivation being to create better software and make everyone's life a bit better*" [SM28].

These participants affirmed their commitment to the communities and values of free software in their contribution decisions. Still, some participants clarified that they were more nuanced in evaluating projects: their decisions were informed by how the contribution experience felt or based on an established history between the corporation and FLOSS communities. Indeed, several participants described the importance of familiarity with a company's reputation in FLOSS communities, as this better enabled the assessment of risk for participation decisions. Even in situationally-informed decisions, contributors expressed a limit to such nuance as some entities were to be avoided at all costs: *"I [am] sensitive to a small set of organizations with whom I do not want any association"* [SM21]. Each of these risk and goal factors leads some FLOSS contributors to exhibit hesitancy in joining or to completely avoid corporate-backed projects.

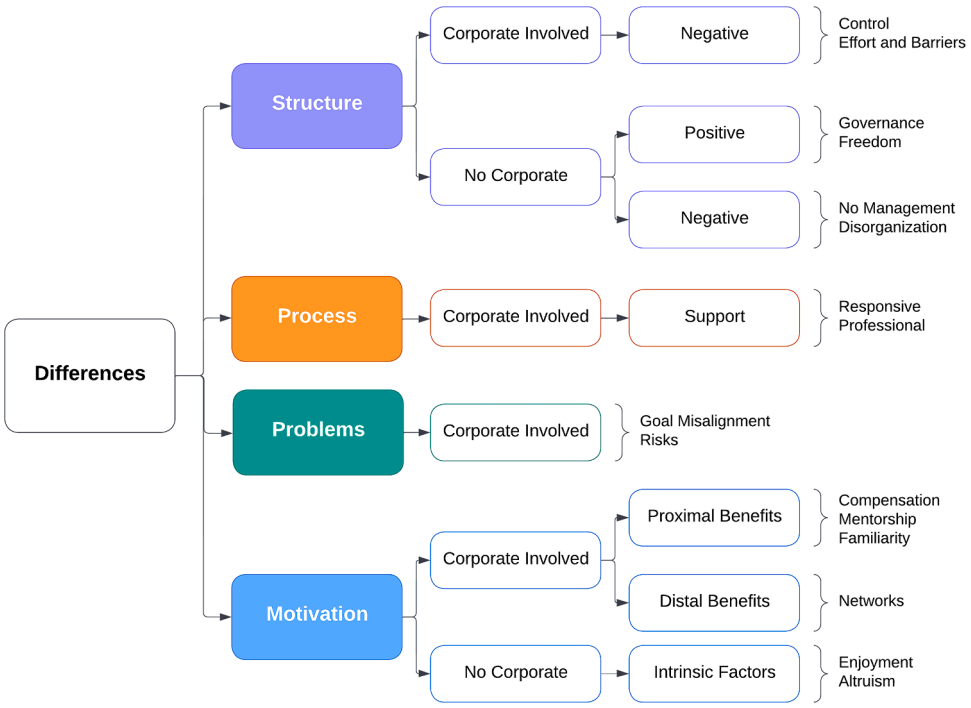


Fig. 3. Themes and subthemes in participant responses regarding differences between projects with and without corporate involvement.

5.3. Survey research: corporations and diversity

With respect to **RQ1c**, participants generally held neutral or positive views of corporate effects on diversity in FLOSS projects (Figs. 4 and 5). Only a very small proportion of participants in North America felt that corporations were seriously detrimental to social diversity in FLOSS projects (Fig. 5). Perceptions of participants across gender and geographical groups regarding effects of diversity were thus somewhat more positively skewed than perceptions of corporations in FLOSS development generally.

The results of this survey research provide evidence for the effect of corporate activities and decisions on individual decision making related to participation in FLOSS development. Participants in the sample not only confirmed that there are key differences in individual preferences, experiences, and beliefs with respect to project types, they provided a richer understanding of their perceptions through descriptions of the rationale which informs their perspective and in turn guides their decision making. We next describe the results of a complementary approach to our survey research: an analysis of big data extracted from FLOSS project ecosystems in GitHub.

5.4. GitHub projects research

5.4.1. Sustained participation

Relevant to **RQ2a**, we provide summary statistics for contributor tenure in the three levels of project type and describe the results of the survival analysis conducted to examine differences in sustained participation between fully corporate and non-corporate projects. The average tenure is similar between fully corporate and non-corporate projects, but lower in projects which receive corporate support (Table 8). Differences in survival probabilities between project types are provided in Table 9 and visualized in Fig. 6.

The largest differences in survival probabilities for corporate and non-corporate projects occurs in the first three months of contribution. During this first quarter of participation, contributors in projects with no corporate influence have a 13% greater chance of remaining than contributors in projects owned and maintained by corporations. This difference decreases after the first year, with contributors in both project types having similar likelihood of sustained participation. Stated another way, the risk for leaving is lower at the beginning for non-corporate projects—the curve begins at approximately 70% chance of survival—and this risk increases to a similar level observed in fully corporate projects after a year. To determine if the difference between groups is statistically significant, a log-rank test was conducted. The results of this test indicate that there is a significant difference in overall survival between projects with and without corporate involvement ($X^2 = 10.70$, $p = .001$). We thus reject the null hypothesis.

5.4.2. Gender and geographic diversity

For gender diversity, we report two Blau index values based on genderComputer labeling and pronoun-enhanced genderComputer labeling. Density plots are provided in Fig. 7. Relevant to **RQ2b-i**, gender diversity is generally low in projects (<0.3). A Welch's unequal variances t-test, a modification of the two sample t-test that is appropriate for cases in which there are unequal groups and there is unequal variance between those groups, was applied to the set of Blau index values for gender diversity. We report No significant difference was found between the two project types of interest with the genderComputer labeling, $t(22) = 0.74$, $p = .47$, $d = 0.28$ (lower CI: -0.50; upper CI: 1.06), with no corporate influence projects ($M = 0.16$, $SD = 0.07$) having, on average, similar levels of gender diversity as company-owned projects ($M = 0.14$, $SD = 0.09$). There was also no significant difference between the two project types of interest with the pronoun-enhanced genderComputer labeling, $t(22) = 0.94$, $p = .36$, $d = 0.33$ (lower CI: -0.45; upper CI: 1.11), with no corporate influence projects ($M = 0.18$, $SD = 0.09$) having, on average, similar levels of gender diversity as company-owned projects ($M = 0.16$, $SD = 0.04$). The enhancement of genderComputer labels using pronoun information somewhat alters calculated diversity levels as indicated by the plots. This is due in part to the fact that a third group has been added in the calculation of the Blau index for gender diversity which in turn increases the maximum possible value. The project with the highest gender diversity across both calculations of the Blau index falls into the no corporate influence category. When considering the top five projects, however, the majority shifts from company-owned to no corporate influence with pronoun-enhanced gender resolution. While the diversity values change and increase with

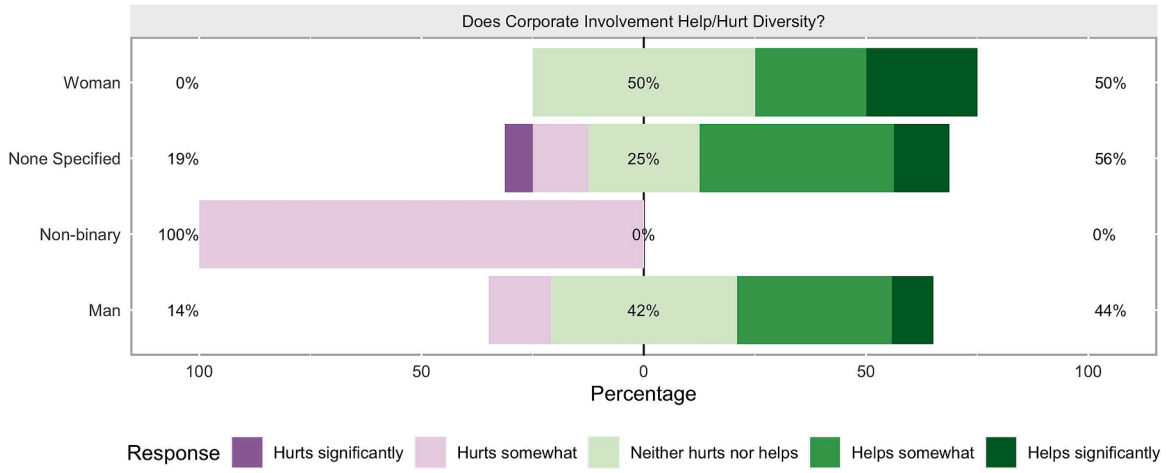


Fig. 4. Gender differences in the perceived effect of corporate involvement on social diversity in FLOSS development.

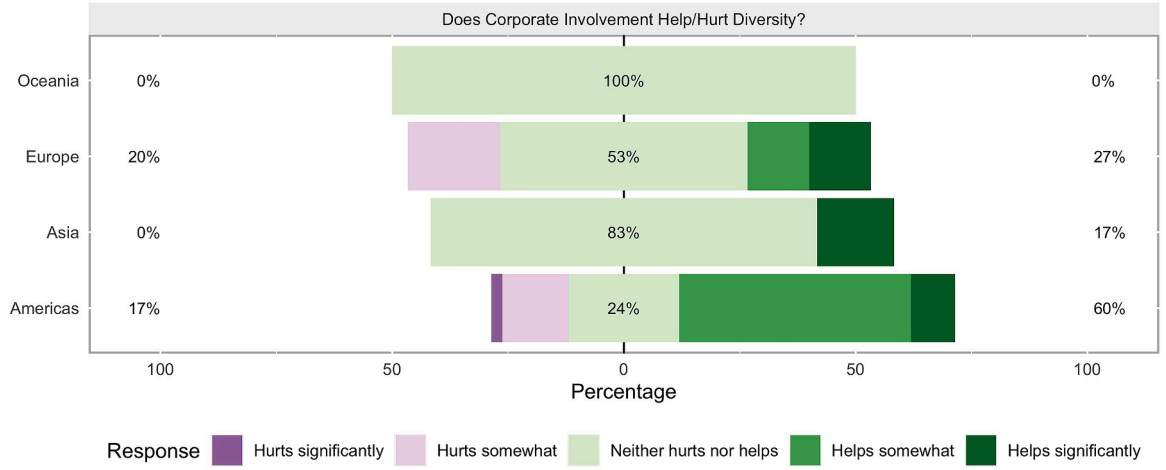


Fig. 5. Geographic differences in the perceived effect of corporate involvement on social diversity in FLOSS development.

Table 8
Summary statistics for project tenure aggregated at project type level.

Project Type	Min	Max	Median	Mean (SD)
Non-Corporate	1	54	10	13.37 (±12.74)
Receive Corporate Support	1	57	1	9.60 (±12.98)
Fully Corporate	1	56	9	13.79 (±14.31)

Table 9
Survival probabilities by project type at two time points. Survival probabilities at these time points represent the greatest difference (quarter 1) and convergence (quarter 12) between project types.

Quarter	Project Type	Probability	SE	Lower, Upper 95% CI
1	Non-Corp.	0.71	0.01	0.70, 0.73
	Corporate	0.58	0.01	0.55, 0.66
12	Non-Corp.	0.45	0.01	0.43, 0.47
	Corporate	0.46	0.01	0.44, 0.48

pronoun enhancement, the difference between project types remains non-significant.

Relevant to **RQ2b-ii**, country diversity is relatively high (>0.75). The density plot is provided in Fig. 8 and summary statistics in Table 10. The skewness and prevalence of ties between project types limited the application of statistical tests to country diversity data. We examined country diversity between project types within a project ecosystem and observed that six out of nine (67%) projects with the highest country diversity in their ecosystems had no corporate involvement. Additionally, compared to the other two types of projects, projects owned by a firm have the lowest minimum country diversity and a lower maximum country diversity. While this does not provide unequivocal evidence that projects with no corporate involvement tend to be more globally open, it does suggest that there may be important differences in country diversity between project types. In sum, no project types are more or less open to different gender groups and projects with no corporate involvement may be more open to globally-distributed groups.

6. Discussion

The goal of this research was to characterize the perceptions of diversity and corporations held by FLOSS project contributors and to examine if and how this is reflected in observable behavior in GitHub, a social coding platform. Relevant to our first research question, the results of the survey highlight the heterogeneity of FLOSS contributors' beliefs, behaviors, and experiences associated with participation and corporations. Specifically, we find that:

RQ1a Contributors report experiencing important differences between projects based on corporate involvement, specifically with respect to structure, processes, goals, risk, values, and motivation.

RQ1b Contributors vary in the degree to which their participation decisions are influenced by the presence of corporations, with some

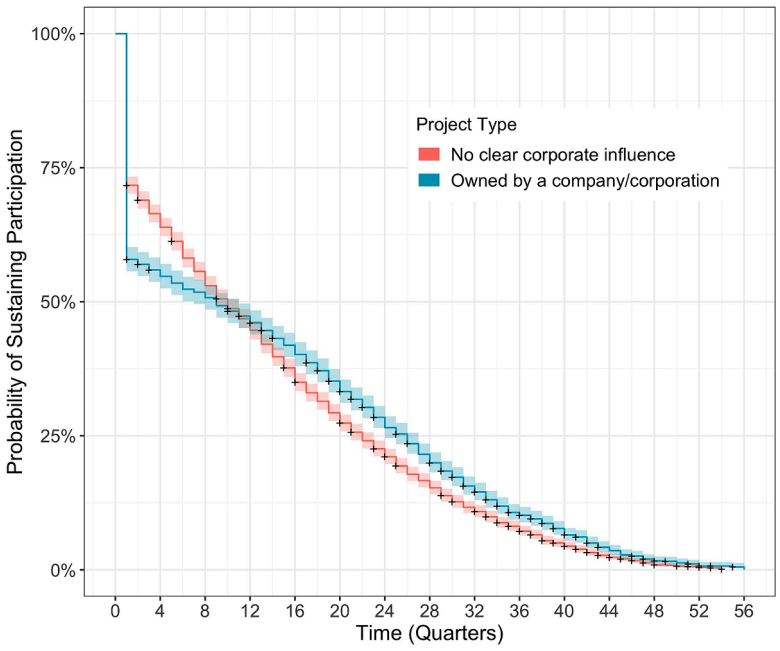


Fig. 6. Survival curves for contributors to projects with and without corporate involvement. The probability of maintaining activity in a project is plotted on the *y*-axis. Time, aggregated in 3-month periods (i.e., quarters), is plotted on the *x*-axis.

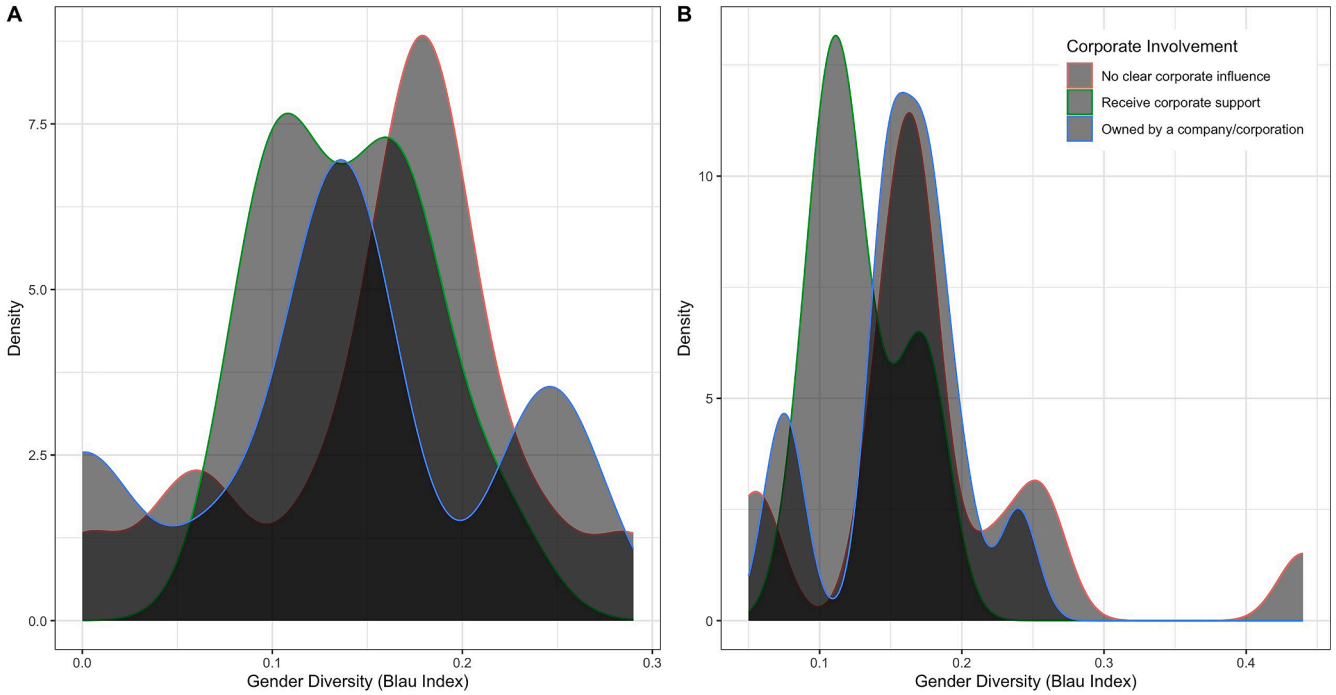


Fig. 7. The distribution of gender diversity by project type, visualized with density plots. In plot A, contributors were labeled using the genderComputer tool (left). In plot B, contributors were labeled using the genderComputer tool with enhancement from specified pronouns (right).

having very strong anti-corporation stances and others being drawn to potential personal benefits associated with corporate engagement.

RQ1c Contributors also vary in the degree to which they perceive corporations as being helpful or harmful to the achievement of diversity, but many assume the effect is nonexistent or minimally positive.

The results of the GitHub data analysis indicate that there are some differences in participation associated with corporate engagement, specifically with respect to early commitment to a project. Specifically, we observe that:

RQ2a Contributors to FLOSS projects owned and managed entirely by firms are less likely to sustain participation in the first quarter of participation when compared to projects without corporate engagement.

RQ2b In general, there is no clear evidence that differences in social diversity are associated with corporate involvement in FLOSS development.

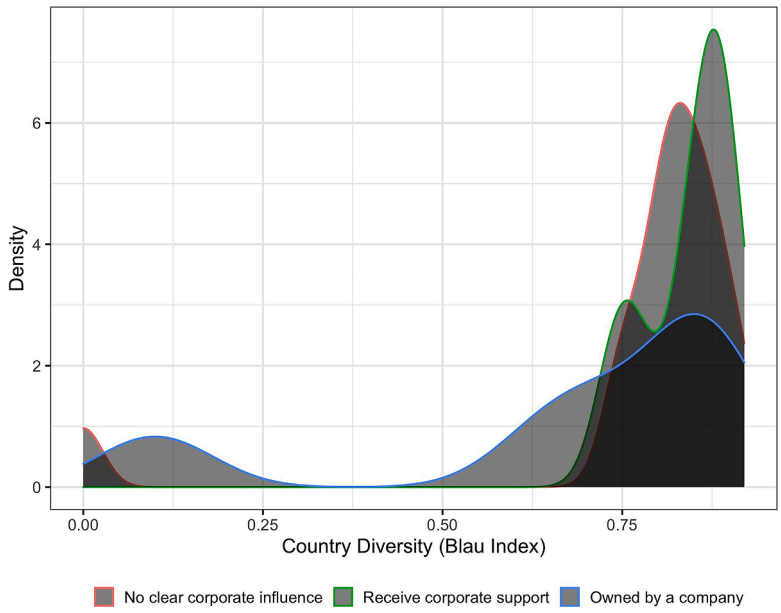


Fig. 8. The distribution of country diversity by project type, visualized with density plots.

Table 10
Summary statistics for country diversity by project types as measured using the Blau index.

Statistic	No Corp. Involvement	Corp. Support	Corp. Owned
Minimum	0.44	0.77	0.07
Median	0.82	0.86	0.84
Maximum	0.92	0.92	0.89
Mean	0.80	0.85	0.68

i. FLOSS projects owned and managed entirely by firms are in general neither significantly better nor worse at forming gender diverse groups when compared to projects without corporate engagement.
ii. There is some evidence that FLOSS projects owned and man-aged entirely by firms may be less likely to form culturally diverse groups when compared to projects without corporate engagement.

A key takeaway for this research is that FLOSS contributors experience important differences between projects based on the presence of corporations in them. When looking at behavioral data extracted from FLOSS projects, we see that the presence of a corporation or company is not associated with greater levels of diversity, even though some contributors perceive corporations as having a positive influence on this dimension of group composition. Organizations interested in affecting change in FLOSS projects must reflect on why their own internal practices, norms, and/or culture have not already produced diverse collaborations. Meaningful action and policy relevant to internal dynamics is paramount before any attempt to impose top-down measures on FLOSS communities, which display sensitivity to control, is made in projects. This is essential for not only influencing those communities, but also ensuring the integrity and longevity of diversity as a commitment.

Communities of contributors who are especially oriented towards software freedom rather than personal benefits and profit are unlikely to respond positively when they perceive organizational actions as undue or unjustified control. Surface-level actions without meaningful change to organizational policy are seen as empty gestures centered on presenting an image of inclusivity and concern for equity, rather than an embodiment of such values. As one survey participant expressed, *"I'm displeased by GitHub's move to rename the "master" branch, I think it's a bullshit statement and it doesn't solve any of our modern-day problems"* [GH11]. Research on diversity and anti-discrimination programs in the workplace have shown that such initiatives, motivated by a desire to protect public image and reduce legal risks, do not only fail to increase diversity, they effectively reduce it further (Dobbin and Kalev, 2022).

The results of the GitHub data analysis suggest that company-owned projects tend to have lower levels of country diversity. Even though country diversity was generally high across all project types, it was limited in certain continents and geographical regions—Africa, Oceania, West and South Asia, and South America had lower levels of representation compared to North America, Europe, and East Asia. Organizations interested in extending use of FLOSS and participation in projects to a broader group may then want to consider ways that they can lower barriers to contribution for these underrepresented populations. A survey participant provided an example of a relevant barrier that limits the achievement of diversity:

> "*While open source is used worldwide, not everyone has the technical ability or time to contribute. Being able to measure diversity in projects, and work to improve it, would be a huge benefit to the broader open source community. For example, much code is written by English-speaking programmers, making it challenging for non-native speakers to follow comments, method / function names, etc.*" [SM34].

Addressing such barriers can then help move FLOSS towards greater inclusivity at the global level. In this specific case, organizations can intentionally devote resources to translation, allocating funds to services and paying non-English speakers in underrepresented countries to support the completion of this work. Such an action would not only immediately confer benefits to global labor, it could potentially have cascading effects by providing access to a much broader audience.

The results of the RTA applied to participant responses regarding their experiences and feelings towards corporations in FLOSS development reveal that underlying motivational and reward differences for working in varied contexts. In the first case, a subset of FLOSS contributors delineate the role of their economic needs and management-style preferences. These participants note that their contributions to projects with corporate engagement grant them immediate and more distal benefits in the form of compensation and networks, respectively. They additionally note that the management styles employed in projects with corporate involvement ensure that they have access to personnel and resources to enable the completion of their work.

In the second case, a subset of FLOSS contributors describe the impediments of corporate management style to their work, making it

needlessly laborious through increased bureaucracy and constraining their ability to innovate due to top-down control. Further, they may view the profit-based goals of the corporation to not be in line with their personal values—improving codebases as a greater good—or the broader values of openness and unrestricted use in the free software development paradigm. This extends prior work in the organizational sciences by elaborating the specific economic, organizational, and work-related factors that drive observed membership change in technology development, and provides evidence for a set of FLOSS contributor types that can be differentiated along these three dimensions.

The importance of these dimensions and contributor types is more salient when connecting them to the development of interventions to improve diversity and inclusion in FLOSS. The former group appears amenable to top-down interventions under conditions where they are able to continue to derive economic benefits and resources to support this work remain available. The latter group in contrast may reject top-down interventions, particularly when they lack transparency and leadership fails to make explicit how such interventions will produce benefits for code quality. As such, we contend that developer and community stance towards corporate engagement, which exists on a spectrum from anti to pro, is an important dimension for firms to evaluate when considering top-down mechanisms for increasing inclusivity in FLOSS projects. The findings presented here can inform future research to test and develop interventions in FLOSS in order to make advances in this important contemporary form of work.

The adoption of FLOSS development practices by organizations and the growing number of collaborative multi-organizational/institutional collaborative initiatives may indeed be useful in encouraging a more inclusive, global workforce. Yet, organizations invested in making change must practice care in pursuing such collaborations. In particular, the responses of survey participants suggest that some corporations lack social capital with respect to FLOSS communities (contributors and users). Therefore, it is important for organizations to be judicious in forming collaborations and to maintain the trust of those they choose to collaborate with over time. Further, they should heed the advice given to those seeking to participate in FLOSS development: thoroughly vet the history of potential collaborators with affected communities.

Acquisitions in FLOSS ecosystems can also be quite disruptive depending on the social-normative attitudes held by the FLOSS community and may produce a loss with respect to advances made towards greater inclusivity and diversity. For example, Microsoft has a contentious relationship with FLOSS communities, being an early opponent of free and open source licensing. Following Microsoft's acquisition of GitHub, some survey participants shared that they migrated away from the platform, while others braced for an inevitable departure, taking a 'wait and see' attitude as to when corporate ownership would impinge on their values.

6.1. Limitations

6.1.1. Qualitative survey

Our study has limitations that should be acknowledged and may be addressed in future research. The survey data collected in the study is somewhat affected by sampling bias due to the open invitation approach employed in the first phase of distribution. Invitations in the second phase of survey distribution were based on an empirically-driven sampling approach which served to minimize this bias. Related to the sample, the demographics of the survey participants are relatively representative of the study population but do not necessarily capture the experiences of underrepresented groups in FLOSS development. Future work can aim to recruit FLOSS contributors from underrepresented groups more directly, including marginalized genders and the Global South.

6.1.2. Quantitative project analysis

Although our sample was carefully selected using a coverage algorithm that maximizes representativeness, our final sample may not have included every possible scenario for corporate involvement. It was not feasible for us to qualitatively label all projects in the GHTorrent data dump according to level of corporate involvement. The distribution of project information across the internet limits this labeling. For the final sample, we focused on labeling projects according to their status at the time of data collection. While this enabled us to form distinct categories based on project involvement, it did not take into account changes to status over time. For example, one project in the receives corporate support category was initially owned by a company but shifted to release of company ownership and only received support when we collected the data. Nonetheless, the corporate involvement categories constructed in this research were associated with statistically significant differences in sustained participation. This limitation also points to opportunities for future research to develop automated labeling (e.g., with natural language processing, language models, and web scraping) and examine status changes over time to represent the full spectrum of corporate involvement.

The gender of contributors in the sampled GitHub projects was inferred through limited information provided in user profiles. In March of 2023, GitHub announced that users could add pronouns to their profile[6]. Such user-provided information can potentially be leveraged to enhance classification in studies of gender and group diversity. The utility of this feature for research on gender is constrained by the extent to which users choose to adopt it. In this study's sample, only 7% of GitHub contributors chose to add pronouns to their profile. Still, the enhancement of gender labeling afforded by pronouns produces some differences in observed diversity levels as it enables analysis beyond a gender binary. Ultimately though, pronouns, like names, are informative but uncertain proxies for gender. For example, a non-binary individual may choose to use a set of gendered pronouns (e.g., she/her), or a blend of gendered pronouns (e.g., both he/him and she/her) with or without gender-neutral pronouns. Pronoun information can only be accessed with an authenticated account to protect the privacy of users[7]. Greater adoption of pronoun use by platform users may serve to improve the precision and representation of gender in future research, but privacy protections should be considered if data are to be publicly shared.

7. Conclusion

We examined differences in participation diversity between projects with different organizational structures and control mechanisms, showing that they vary in terms of both perceived and actual openness and inclusivity. The qualitative component of this research revealed subgroups among FLOSS contributors in terms of values and goals: study participants report that they are guided by a desire to reap personal benefits or maintain codebases as a social good. In other words, FLOSS contributors may potentially be differentiated on the basis of their orientation toward personal gain or collective good. Prior research on diversity has examined individualism and collectivism on the basis of geographically-based culture differences and its association with behavior in online programming communities (Zolduoarrati et al., 2022). Our work provides understanding of an orthogonal dimension to consider as members of individualistic cultures can be oriented towards collective good or personal gain. The behavior of developers observed in platforms may thus reflect the interaction of individual orientation (personal versus collective) and cultural orientation (individualistic or collectivist).

The results of the quantitative component research suggest that variations in participation likely occur on the basis of corporate involvement. While there was no evidence of differences along the

[6] https://github.blog/changelog/2023-03-01-add-pronouns-to-your-github-profile/

[7] https://github.com/orgs/community/discussions/78811

dimension of gender, some evidence was found that projects with no corporate influence exhibit higher levels of openness along the dimension of geographical location, as they had among the highest levels of country diversity. Future research can investigate how recruitment and hiring policies may influence such differences in geographic participation. Altogether, this study suggests that to differentially influence participation and both encourage and maintain diversity in FLOSS projects, it is important to develop strategies that are appropriate for the stances and orientations of the contributor base.

CRediT authorship contribution statement

Olivia B. Newton: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Stephen M. Fiore:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A link to an OSF project containing study materials has been provided.

References

Diversity, inclusion, and Belonging at GitHub in 2022, 2022. https://github.com/about/diversity/report.

Barcomb, A., Kaufmann, A., Riehle, D., Stol, K.-J., Fitzgerald, B., 2018. Uncovering the periphery: a qualitative survey of episodic volunteering in free/libre and open source software communities. IEEE Trans. Softw. Eng. 18. https://doi.org/10.1109/TSE.2018.2872713.

Birkinbine, B.J., 2015. Conflict in the commons: towards a political economy of corporate involvement in free and open source software. Polit. Econ. Commun. 2 (22).

Blincoe, K., Harrison, F., Damian, D., 2015. Ecosystems in GitHub and a method for ecosystem identification using reference coupling. In: Proceedings of the 12th Working Conference on Mining Software. IEEE Press, Piscataway, NJ, pp. 202–211. https://doi.org/10.1109/MSR.2015.26.

Braun, V., Clarke, V., 2019. Reflecting on reflexive thematic analysis. Qual. Res. Sport Exerc. Health 11 (4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806.

J. Bryer, K. Speerschneider, likert: analysis and visualization likert items (2022). https://github.com/jbryer/likert.

Burr, V., 2015. Social Constructionism, 3rd Edition. Routledge, London. https://doi.org/10.4324/9781315715421.

Byrne, D., 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. Qual. Quant. 56 (3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y.

Cambon, J., Hernangómez, D., Belanger, C., Possenriede, D., 2021. tidygeocoder: an r package for geocoding. J. Open Source Softw. 6 (65), 3544. https://doi.org/10.21105/joss.03544.

Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G., 2003. Survival analysis part i: basic concepts and first analyses. Br. J. Cancer 89 (2), 232–238. https://doi.org/10.1038/sj.bjc.6601118.

Clarke, V., Braun, V., 2013. Successful Qualitative Research: A Practical Guide for Beginners. Sage Publications.

Clarke, V., Braun, V., 2014. Thematic analysis. In: APA Handbook of Research Methods in Psychology, 3, pp. 1947–1952. https://doi.org/10.1037/13620-004.

Coleman, E.G., 2013. Coding Freedom: The Ethics and Aesthetics of Hacking. Princeton University Press, Princeton.

Crenshaw, K., 2018. Demarginalizing the Intersection of Race and Sex: A black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics, 1st Edition. Routledge, pp. 57–80. https://doi.org/10.4324/9780429500480-5.

Dahlander, L., Magnusson, M.G., 2005. Relationships between open source software companies and communities: observations from Nordic firms. Res. Policy 34 (4), 481–493. https://doi.org/10.1016/j.respol.2005. 02.003.

David, P.A., Shapiro, J.S., 2008. Community-based production of open-source software: what do we know about the developers who participate? Inf. Econ. Policy 20 (4), 364–398. https://doi.org/10.1016/j.infoecopol.2008.10.001.

Dobbin, F., Kalev, A., 2022. Getting to Diversity: What Works and What Doesn't. The Belknap Press of Harvard University Press, Cambridge, Massachusetts.

El Asri, I., Kerzazi, N., 2019. Where are Females in OSS Projects? Socio Technical Interactions, 568. Springer International Publishing, Cham, pp. 308–319. https://doi.org/10.1007/978-3-030-28464-0_27.

Fitzgerald, B., 2006. The transformation of open source software. MIS Q. 30 (3), 587. https://doi.org/10.2307/25148740.

Fronchetti, F., Wiese, I., Pinto, G., Steinmacher, I., 2019. What attracts newcomers to onboard on OSS projects? TL;DR: popularity. In: Vol. 556 of IFIP Advances in Information and Communication Technology. Springer, Cham, pp. 91–103. https://doi.org/10.1007/978-3-030-20883-7_9.

Germonprez, M., Allen, J.P., Warner, B., Hill, J., McClements, G., 2013. Open source communities of competitors. Interactions 20 (6), 54–59. https://doi.org/10.1145/2527191.

Germonprez, M., Kendall, J.E., Kendall, K.E., Mathiassen, L., Young, B., Warner, B., 2017. A theory of responsive design: a field study of corporate engagement with open source communities. Inf. Syst. Res. 28 (1), 64–83. https://doi.org/10.1287/isre.2016.0662.

Germonprez, M., Lipps, J., Goggins, S., 2019. The rising tide: open source's steady transformation. First Monday. https://doi.org/10.5210/fm.v24i8.9297.

Gijbels, I., 2010. Censored data. WIREs Comput. Stat. 2 (2), 178–188. https://doi.org/10.1002/wics.80.

Golzadeh, M., Decan, A., Legay, D., Mens, T., 2021. A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments. J. Syst. Softw. 175, 110911 https://doi.org/10.1016/j.jss.2021.110911 arXiv: 2010.03303.

Gralha, C., Goulao, M., Araujo, J., 2019. Analysing gender differences in building social goal models: a quasi-experiment. In: 2019 IEEE 27th International Requirements Engineering Conference (RE). IEEE, Jeju Island, Korea (South), pp. 165–176. https://doi.org/10.1109/RE.2019.00027.

Haraway, D., 1988. Situated knowledges: the science question in feminism and the privilege of partial perspective. Feminist Stud. 14 (3), 575. https://doi.org/10.2307/3178066.

Harrison, D.A., Klein, K.J., 2007. What's the difference? diversity constructs as separation, variety, or disparity in organizations. Acad. Manage. Rev. 32 (4), 1199–1228. https://doi.org/10.5465/amr.2007.26586096.

Hilderbrand, C., Perdriau, C., Letaw, L., Emard, J., Steine-Hanson, Z., Burnett, M., Sarma, A., 2020. Engineering gender-inclusivity into software: ten teams' tales from the trenches. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. ACM, Seoul South Korea, pp. 433–444. https://doi.org/10.1145/3377811.3380371.

M. Horikoshi, Y. Tang, ggfortify: data visualization tools for statistical analysis results (2018). https://CRAN.R-project.org/package=ggfortify.

Imtiaz, N., Middleton, J., Chakraborty, J., Robson, N., Bai, G., MurphyHill, E., 2019. Investigating the effects of gender bias on GitHub. In: Proceedings of the 41st International Conference on Software Engineering. IEEE Press, Montreal, QC, Canada, pp. 700–711. https://doi.org/10.1109/ICSE. 2019.00079.

Jarczyk, O., Jaroszewicz, S., Wierzbicki, A., Pawlak, K., JankowskiLorek, M., 2018. Surgical teams on GitHub: modeling performance of GitHub project development processes. Inf. Softw. Technol. 100, 32–46. https://doi.org/10.1016/j.infsof.2018.03.010.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53 (282), 457–481. https://doi.org/10.1080/01621459.1958.10501452.

King, D.K., 1988. Multiple jeopardy, multiple consciousness: the context of a black feminist ideology. Signs 14 (1), 42–72. https://doi.org/10.1086/494491.

Liu, X., 2012. Survival Analysis: Models and Applications. John Wiley Sons google-Books-ID: bEZpGtw39qgC.

McNicol, A., 2013. None of Your Business? Analyzing the Legitimacy and Effects of Gendering Social Spaces Through System Design. Institute of Network Cultures, Amsterdam, pp. 200–219.

Nadri, R., Rodriguez Perez, G., Nagappan, M., 2021. Insights into nonmerged pull requests in GitHub: is there evidence of bias based on perceptible race? IEEE Softw. 38 (2), 51–57. https://doi.org/10.1109/MS.2020.3036758.

Nafus, D., 2012. patches don't have gender': what is not open in open source software. New. Media Soc. 14 (4), 669–683. https://doi.org/10.1177/1461444811422887.

Nagappan, M., Zimmermann, T., Bird, C., 2013. Diversity in software engineering research. In: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2013. ACM Press, Saint Petersburg, Russia, p. 466. https://doi.org/10.1145/2491411.2491415.

Newton, O.B., 2020. Defining and promoting societal benefits in open source software development. In: GROUP4GOOD Workshop at the 2020 ACM International Conference on Supporting Group Work (GROUP). ACM, Sanibel Island, Florida, USA. https://doi.org/10.17605/OSF.IO/843VS.

O.B. Newton, J. Song, Modeling gender differences in membership change in open source software projects (arXiv:2206.08485) (2022). http://arxiv.org/abs/2206.08485.

Qiu, H.S., Nolte, A., Brown, A., Serebrenik, A., Vasilescu, B., 2019. Going farther together: the impact of social capital on sustained participation in open source. In: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). IEEE, Montreal, QC, Canada, pp. 688–699. https://doi.org/10.1109/ICSE.2019.00078.

Rajanen, M., Iivari, N., 2015. Power, empowerment and open source usability. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. ACM Press, Seoul, Republic of Korea, pp. 3413–3422. https://doi.org/10.1145/2702123.2702441. http://dl.acm.org/citation.cfm?doid=2702123.2702441.

T. Saunders, How many countries are there in 2024? (2024). https://www.sciencefocus.com/planet-earth/how-many-countries-are-there.
Schmaus, W., 2015. Determinism: Social and Economic. Elsevier, pp. 241–246. https://doi.org/10.1016/B978-0-08-097086-8.03126-3.
Sen, R., Nelson, M., Subramaniam, C., 2015. Application of survival model to understand open source software release. Pac. Asia J. Assoc. Inf. Syst. 7 (2) https://doi.org/10.17705/1pais.07201. https://aisel.aisnet.org/pajais/vol7/iss2/1.
Simpson, E.H., 1949. Measurement of diversity. Nature 163 (4148), 688. https://doi.org/10.1038/163688a0, 688.
Tang, Y., Horikoshi, M., Li, W., 2016. ggfortify: unified interface to visualize statistical result of popular r packages. R. J. 8 (2), 474–485. https://doi.org/10.32614/RJ-2016-060.
Tennekes, M., 2018. tmap: thematic maps in r. J. Stat. Softw. 84 (6), 1–39. https://doi.org/10.18637/jss.
Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C., Stallings, J., 2017. Gender differences and bias in open source: pull request acceptance of women versus men. PeerJ Comput. Sci. 3, e111. https://doi.org/10.7717/peerj-cs.111.
T.M. Therneau, A package for survival analysis in r (2024). https://CRAN.R-project.org/package=survival.
Vasilescu, B., Filkov, V., Serebrenik, A., 2015. Perceptions of diversity on GitHub: a user survey. In: 2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering. IEEE, Florence, Italy, pp. 50–56. https://doi.org/10.1109/CHASE.2015.14.
Vasilescu, B., Posnett, D., Ray, B., van den Brand, M.G., Serebrenik, A., Devanbu, P., Filkov, V., 2015. Gender and tenure diversity in GitHub teams. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15. ACM Press, Seoul, Republic of Korea, pp. 3789–3798. https://doi.org/10.1145/2702123.2702549.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., Fran¸cois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.
Winchester, H., Boyd, A.E., Johnson, B., 2022. An exploration of intersectionality in software development and use. In: 2022 IEEE/ACM 3rd International Workshop on Gender Equality, Diversity and Inclusion in Software Engineering (GEICSE). ACM, Pittsburgh, PA, USA, pp. 67–70. https://doi.org/10.1145/3524501.3527605.
Xu, Y., Pace, S., Kim, J., Iachini, A., King, L.B., Harrison, T., DeHart, D., Levkoff, S.E., Browne, T.A., Lewis, A.A., Kunz, G.M., Reitmeier, M., Utter, R.K., Simone, M., 2022. Threats to online surveys: recognizing, detecting, and preventing survey bots. Soc. Work Res. svac023. https://doi.org/10.1093/swr/svac023.
Zolduoarrati, E., Licorish, S.A., Stanger, N., 2022. Impact of individualism and collectivism cultural profiles on the behaviour of software developers: a study of stack overflow. J. Syst. Softw. 192, 111427 https://doi.org/10.1016/j.jss.2022.111427.

Olivia B. Newton, Ph.D. Olivia is a Postdoctoral Scholar in the Institute for Simulation and Training at the University of Central Florida. She is interested in modeling human and technical aspects of collaboration in science and engineering domains to enhance team effectiveness, accelerate innovation, and promote equity in computational technologies.

Stephen M. Fiore, Ph.D. Stephen is Director, Cognitive Sciences Laboratory, and Professor with the University of Central Florida's Cognitive Sciences Program in the Department of Philosophy and Institute for Simulation and Training. He maintains a multidisciplinary research interest that incorporates aspects of the cognitive, social, organizational, and computational sciences in the investigation of learning and performance in individuals and teams. His primary area of research is the interdisciplinary study of complex collaborative cognition and the understanding of how humans interact socially and with technology.