Journal of Software Engineering Research and Development, 2019, 13:3, doi: 10.5753/jserd.2025.4142

Characterizing the Participation of the LGBTQIA+ Community on GitHub

Erick Paiva [**PUC-MG** | ***erick.paiva@sga.pucminas.br***]
Guilherme Carvalho [**PUC-MG** | ***blcpinto@sga.pucminas.br***]
João Pedro Mayrink [**PUC-MG** | ***joao.jesus.1302911@sga.pucminas.br***]
Maria Clara Maruch [**PUC-MG** | ***maria.jabali@sga.pucminas.br***]
Pedro Felix [**PUC-MG** | ***pedro.costa.1298438@sga.pucminas.br***]
Gabriel Pacheco [**PUC-MG** | ***gacampacheco@gmail.com***]
Laerte Xavier [**PUC-MG** | ***laertexavier@pucminas.br***]

Abstract

Representation and inclusion are ongoing concerns in Software Engineering. To address this issue, previous studies present data and discuss strategies to mitigate inequality in team assembly. However, it is still unclear how these concerns enable the inclusion of developers belonging to minority groups in open-source projects. In particular, little is known about the participation of people from the LGBTQIA+ community in the development of popular GitHub projects. Therefore, this study aims to characterize the participation of this community, based on the collection of 4K user profiles and almost 58K repositories. We also seek to understand the behavior and the interaction of these users within this context, as well as to shed light on the creation of strategies that promote greater inclusion. As a result, it was observed that the technical profile of the users is far from that of the general developer community, despite their active participation in popular JavaScript and Python repositories. The social profile is isolated, with few user interactions, although there is a concentration in specific areas on the platform's map.

Keywords: *Diversity, LGBTQIA+, Repository mining, Profiles, Interaction*

1 Introduction

In software engineering undergraduate courses, individuals who are lesbian, gay, bisexual, transgender, queer, intersex, asexual, and other sexuality and gender identifications (LGBTQIA+) are among the underrepresented groups that often face feelings of exclusion and marginalization (Richard et al., 2022). Even though it is a complex scenario that involves historical, social, and cultural aspects, it is important to note that studies and strategies have been proposed to address inequality over the years (Garcia et al., 2023; Wang and Hejazi Moghadam, 2017).

A prior study investigated project choices on GitHub, focusing on productivity, programming language experience and developers' social connections (Casalnuovo et al., 2015). This research identifies behavioral patterns to promote positive social interactions and support software development. In addition, it can help enrich the discussion of LGBTQIA+ inclusion on GitHub, promoting a more welcoming environment.

GitHub is a platform that makes it possible for developers to share software projects (Garcia et al., 2023). Over time, the platform has established itself as a reference for comprehensive studies on technical and social organization (Vasilescu et al., 2015a). The analysis of available data offers a promising opportunity for researchers to investigate social diversity and its relations with the characteristics of projects (Aué et al., 2016). However, little is known about the LGBTQIA+ community's participation on GitHub, their activities, and their interactions. Therefore, the issue addressed in this research is the scarcity of specific studies that analyze and comprehend the contribution of these people to the GitHub.

In this context, the motivation for conducting this research is to achieve a deeper understanding of the LGBTQIA+ community on GitHub, as current literature lacks comprehensive studies in this area. Thus, this study is justified by the fact that organizations and repository owners can use the information obtained to make strategic decisions about how to attract more members of this community to their projects. This helps those responsible to implement effective strategies to build a welcoming and inclusive environment for all contributors. Furthermore, by characterizing the participation of this population, this study can inspire other research and initiatives that promote equal opportunities and the representation of minorities in the technology sector.

Particularly, the general objective of this study is to characterize the participation of the LGBTQIA+ population on the GitHub platform, presenting a general overview of this group. To achieve this goal, the following research questions (RQs) are proposed[1]:

- RQ.1 What is the technical profile of GitHub users belonging to the LGBTQIA+ community?
- RQ.2 What are the characteristics of repositories to which GitHub users belonging to the LGBTQIA+ community contribute?
- RQ.3 What is the social profile of GitHub users belonging to the LGBTQIA+ community amongst themselves?

[1]We acknowledge that this paper is an extension from a previous work (Paiva et al., 2023), in which we first characterized the community itself. In this extension paper, we strengthen our results by including two new RQs, in bold. They aim to deepen the initial characterization and compare the LGBTQIA+ community against other GitHub contributors.

- RQ.4 What are the sentiments related to comments on pull requests made by the LGBTQIA+ population on GitHub?
- RQ.5 What is the difference in the technical profile of people belonging to the LGBTQIA+ community and those who do not?

The remainder of this paper is organized as follows: Section 3 explores related work, and Section 4 describes the methodology used for the study. Section 5 presents the obtained results. Section 6 discusses the proposed research questions. Section 7 details the threats to validity and their mitigations. Finally, Section 8 presents the conclusion of this study.

2 Background

The participation and representation of marginalized communities in the technology sector, particularly within open-source platforms like GitHub, have been subjects of increasing interest and importance (Vasilescu et al., 2015b). Open-source software (OSS) is a crucial component of the modern software development landscape, fostering collaboration, innovation, and the democratization of technology (Aberdour, 2007). However, the inclusivity of these platforms has often been called into question, with particular concerns regarding the visibility and experiences of minority groups, including the LGBTQIA+ community (Janzen et al., 2018).

2.1 The LGBTQIA+ Community in Tech

Alan Turing, widely regarded as one of the founding figures of modern computing, was also a member of the LGBTQIA+ community and faced significant challenges, including persecution due to his sexual orientation, reflecting the discrimination that many queer individuals in tech still endure today (Wall, 2023). The LGBTQIA+ community, encompassing a wide range of sexual orientations and gender identities, faces unique challenges in the tech industry (Albusays et al., 2021). These challenges include discrimination, lack of representation, and the additional burden of navigating their professional identities in environments that may not always be supportive.

Despite these challenges, LGBTQIA+ individuals contribute significantly to the tech industry, bringing diverse perspectives and fostering innovation (van der Meulen and Revilla, 2008). Broadening participation in computing (BPC) has been a key focus of the National Science Foundation (NSF) for over two decades, addressing the inclusion of historically underrepresented groups, yet gender and sexual diversity remains underexplored in computing education research (DuBow et al., 2024). Acknowledging and addressing these disparities is essential for fostering a more inclusive and innovative tech environment, where all individuals—regardless of identity—can thrive and drive the industry forward.

2.2 Open-Source and Inclusivity

Open-source platforms like GitHub have the potential to level the playing field by providing a merit-based environment where contributions can be made by anyone, regardless of their background (Zhao et al., 2024). However, these platforms are not immune to the biases and exclusionary practices that exist in broader society. There are ongoing concerns about the inclusivity of OSS communities, including issues related to harassment, discrimination, and the recognition of contributions from marginalized groups (Bosu and Sultana, 2019).

Prior research on diversity in OSS has primarily focused on gender and racial diversity, with limited attention given to the LGBTQIA+ community (Sultana et al., 2024). Studies have shown that inclusive practices and diverse teams lead to better software development outcomes, yet there is a gap in understanding how these dynamics play out for LGBTQIA+ individuals (Gila et al., 2014).

3 Related Works

In this section, the studies related to this research are addressed. Therefore, the analyzed works encompass the use of GitHub as a database, user interaction on GitHub, diversity in software engineering, and interest in projects related to gender and the LGBTQIA+ community in computing. Due to the specificity of this work, it is important to highlight the difficulty in articulating the results of this research with the limited number of previous studies.

Initially, profiles involved in open-source projects on GitHub generate content such as code, comments, feature requests, discussions, and bug reports (Almarzouq et al., 2020). The platform allows users to interact in various ways, both as individuals or as organizations. GitHub provides an Application Programming Interface (API) that enables the extraction of interaction data. This interface can be accessed via REST (Representational State Transfer) or GraphQL, though it is necessary to manage the number of requests due to the API's rate limits. In addition, it is possible to filter results using some search criteria in the request. Therefore, the platform provides a wide range of information about projects, users, and organizations, which is useful to collect the necessary information for research. This research focuses on characterizing the participation of the LGBTQIA+ community, leveraging the comprehensive data available from GitHub's API to analyze their contributions and interactions on the platform.

Secondly, developers are observed to tend to choose projects in which they have previous working relationships (Casalnuovo et al., 2015). This study aimed to understand how developers choose new projects to contribute to on GitHub. For this, the research selected developers who had been active for at least 5 years and had made at least 500 commits in ten different repositories. Various factors were analyzed, such as productivity, programming language experience, and the existence of previous social ties between a developer and project members. This study relates to the present work since both seek to analyze and evaluate developer participation on GitHub. However, our research specifically investigates the inclusion and participation patterns

of the LGBTQIA + community, providing information on how social bonds within underrepresented groups influence project selection and participation.

Another related work identifies the diversity of social groups found in research in the field of software engineering (Dutta et al., 2023). The study selects research that uses different social characteristics to analyze the studied groups, identifying them. Data was collected from 79 research papers containing 105 participant studies, identifying a total of 12 diversity categories. Additionally, the work proposes a model for researchers to evaluate the inclusion of social group diversity in their research. This work aims to highlight the existence of content to be studied about a specific social group on the GitHub platform and the diversity of analyses possible with this information about the LGBTQIA+ community. Similarly, our study aims to shed light on the participation of LGBTQIA+ individuals in open-source software development, contributing to the broader understanding of diversity in software engineering.

Moreover, the sentiment analysis about pull requests reveals how developers' emotions fluctuate during the code review process, providing insights into collaboration dynamics and potential areas for improvement (Kumar et al., 2022). The primary objective of this study is to understand the temporal evolution of developer sentiments and their correlations with different programming languages. The authors employ a methodology that involves analyzing comments from commits, pull requests, and issues using sentiment analysis tools like SentiStrength to determine the polarity of emotions expressed. This research is related to this study as both leverage GitHub data to investigate community interactions and sentiment, albeit with different focal points. Insights from the sentiment analysis methodology used by them can be adapted to assist in answering Research Question 4 (RQ4) of the LGBTQIA+ participation study, which examines the sentiment expressed in comments and interactions of LGBTQIA+ developers, enhancing the understanding of their experiences and emotional dynamics on the platform.

Finally, the profiling on GitHub is crucial for understanding how different user groups engage with open-source projects and identifying potential barriers to their participation, as described in Rehman et al. (2020) study. The objective of the research is to track and characterize the contributions of these "newcomer candidates" to understand their participation in OSS projects. The authors employ a mixed-methods approach, combining quantitative and qualitative analysis, to examine whether newcomers practice social coding, the nature of their contributions, the projects they target, and their onboarding success rates. This research is related to this study as both investigate the participation and contributions of specific groups within the GitHub community. While the newcomer study aims to understand the integration and retention of new users, the LGBTQIA+ participation study focuses on characterizing the technical and social profiles of LGBTQIA+ developers. Insights from the newcomer study's methodology can inform the analysis of LGBTQIA+ developers' initial contributions and their integration into OSS projects, enhancing the understanding of their participation dynamics on the platform.

4 Methodology

The proposed study is classified as descriptive research, whose objective is to describe the characteristics of a population or phenomenon. For each RQ, the following metrics were defined with their respective description.

4.1 [RQ.1] What is the technical profile of GitHub users belonging to the LGBTQIA+ community?

The technical profile of users is crucial for understanding the areas of expertise and the type of contributions these users make (Montandon et al., 2021). This helps identify which programming languages and types of projects LGBTQIA+ community members are most active in, providing insight into their skills and technical interests. The metrics used to answer this question are:

M.1 *Most used languages*. To identify the primary programming languages preferred by LGBTQIA+ users, indicating their skills and areas of interest.
M.2 *Frequency of commits*. To measure the activity and ongoing engagement of users in projects.
M.3 *Number of issues and pull requests*. To assess the level of participation and contribution to software development.
M.4 *Account creation date*. To understand the experience and duration of user activity on GitHub.

4.2 [RQ.2] What are the characteristics of repositories to which GitHub users belonging to the LGBTQIA+ community contribute?

Understanding the characteristics of repositories can reveal the types of projects that attract the LGBTQIA+ community and in which types of projects they are most engaged (Saxena and Pedanekar, 2017). This RQ helps delineate the types of projects that receive contributions from the LGBTQIA+ community, identifying if there is a tendency towards certain types of projects or languages. The following metrics help us answer the research question:

M.1 *Number of stars*. To measure the popularity and recognition of the projects they contribute to.
M.2 *Primary language*. To identify the predominant programming languages in the repositories.
M.3 *Closed issues/Total issues*. To evaluate the efficiency and collaboration in projects.
M.4 *Frequency of commits*. To measure the level of activity in projects.
M.5 *Repository creation date*. To determine the maturity of projects.

4.3 [RQ.3] What is the social profile of GitHub users belonging to the LGBTQIA+ community amongst themselves?

The social profile is important for understanding how community members interact with each other, including mutual support, collaboration and sponsoring (Shimada et al., 2022; Wang et al., 2022). This investigation reveals the density and quality of social networks within the community, showing how LGBTQIA+ community members connect and collaborate. For the analysis, these metrics are proposed:

M.1 *Number of followers and followed people from the community*. To measure the level of connection and influence within the community.
M.2 *Number of sponsors and sponsored people from the community*. To evaluate mutual financial support and encouragement.
M.3 *Areas with the most users from the community*. To identify geographical or technological areas with a higher concentration of community members.

4.4 [RQ.4] What are the sentiments related to comments on pull requests made by the LGBTQIA+ population on GitHub?

Analyzing sentiments in comments can provide insights into the social interactions and the emotional environment in which LGBTQIA+ members are involved (Guzman et al., 2014; Lee et al., 2022; Imtiaz et al., 2019). The analysis helps to understand whether members of the community face positive or negative interactions and how these interactions may affect their participation and contribution. Metrics provided to answer this question:

M.1 *Average percentage of positive comments*. To evaluate the level of support and encouragement.
M.2 *Average percentage of negative comments*. To identify the presence of harmful or discriminatory feedback.
M.3 *Average percentage of neutral comments*. To measure neutral and objective interactions.

4.5 [RQ.5] What is the difference in the technical profile of people belonging to the LGBTQIA+ community and those who do not?

Comparing technical profiles can reveal significant differences or similarities, helping to identify any barriers or advantages faced by LGBTQIA+ members (Eaton, 2018; Hu et al., 2018). This characterization allows us to contextualize the contributions and skills of the LGBTQIA+ community in relation to the rest of the GitHub population, offering a comparative perspective. For this, these metrics are proposed:

M.1 *Most used languages*. To compare skills and technological preferences.
M.2 *Closed issues/Total issues of the used repositories*. To compare efficiency and collaboration.
M.3 *Number of issues and pull requests*. To compare the level of participation and contribution.

The remainder of this section describes the procedures (Section 3.1), the methods used for mining GitHub users (Section 4.7), and for collecting repositories (Section 4.8). In addition, the means for assessing users' feelings are described (Section 4.9). Finally, the method to collect data from users outside the community (Section 4.10).

4.6 Procedures

Figure 1 presents an overview of the process adopted in this study, divided into three main steps: collecting data from LGBTQIA+ GitHub users, collecting data from the repositories to which these users contribute, and analyzing and calculating metrics. To do this, Python scripts were created and the data was stored in the MongoDB database. In the first step, requests were made to GitHub's GraphQL API via Python script for an initial collection of users, the data returned by these requests was filtered and stored in the database according to the selection criteria. Next, the second step involved collecting data from the repositories in which these users participate. In this collection, requests were also made to the same API, which collected the participation and contribution data of the users in these repositories, such as pull requests and the programming languages of these pull requests. In addition, the same number of non-participating users from the LGBTQIA+ community and the same data related to these non-participating users were also consulted and stored. Likewise, the repositories where these non-participating users had contributions were also selected and information on their contributions was stored. After collecting the data needed for the analysis, sentiment analysis was carried out on the comments from the pull requests collected and comparative analysis was carried out between the two groups collected.

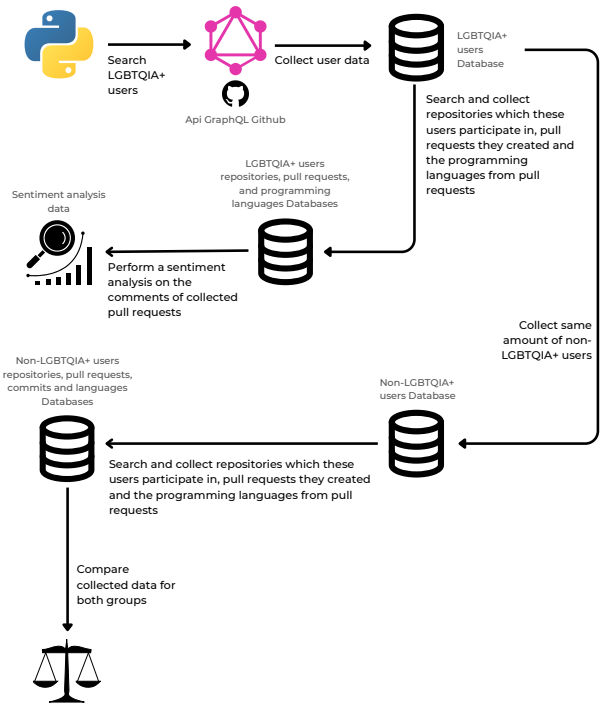

Figure 1. Overview of the adopted methodology

4.7 User Mining

To search for users who self-identify as belonging to the LGBTQIA+ community within the platform, keywords were used as the filter for the name and biography fields in profiles. Thus, users that included the following terms were selected: "queer", "rainbow_flag", "transgender_flag", "nonbinary", "non binary", "lesbian", "bisexual", "asexual", "pansexual", "transgender", "they them", "he them", "she them", "gay", "trans", "transboy", "transgirl", "transwoman" e "transmen".

To validate the users obtained in the first search, a Python script to filter the users was developed to ensure the presence of the chosen keywords. The script checked if the collected user had the keyword in their name or biography independently, rather than being inserted in another word (i.e., profiles that contained these terms as a substring were not selected). In addition, a duplicate check was performed on the list of collected users, since a user could be returned in several queries if they had more than one keyword in their profile. The removal of these duplicates was performed by analyzing the login (i.e., username) of each profile that came up in the query.

During the initial search, the data of all users found was collected to draw their technical profiles. Then, a Python script was developed, using GitHub's GraphQL API, to obtain the commits, pull requests, and programming languages contained in the commits of each pull request made by users (metrics defined for RQ.1). Finally, the rest of the metrics were calculated by identifying the repositories in which these users were active (metrics for RQ.2), and by analyzing the follower list of each of them (metrics for RQ.3). Additionally, metrics for RQ.4 were obtained by analyzing the sentiments related to comments on pull requests made by the LGBTQIA+ population on GitHub. Metrics for RQ.5 were gathered to determine the difference in the technical profile of people belonging to the LGBTQIA+ community and those who do not.

4.8 Collecting Repositories Data

To assess the characteristics of the repositories to which LGBTQIA+ users contribute, the repositories were collected based on the profiles identified in the initial mining. This involved another query using GitHub's GraphQL API, which was used to collect the repositories to which they contributed. The previously collected users' logins were used as identification for this collection. In this process, information such as repository creation date, number of stars, issues, commits, date of the last commit, and main language were extracted.

4.9 Sentiment Analysis

The sentiment analysis performed in this study was conducted using Python's TextBlob[2] library, chosen for its simplicity, speed, and easy integration with the code. The focus was on evaluating the comments associated with pull requests extracted from the database of LGBTQIA+ community users. The collected data was subjected to preprocessing to eliminate irrelevant information or information that could affect the analysis, as well as the removal of links, user mentions, punctuation, stopwords, and tokenization (Etaiwi and Naymat, 2017; Haddi et al., 2013). TextBlob uses a machine learning model to classify the polarity (positive, negative, or neutral) between -1 and 1, whereby the -1 refers to the negative sentiment and the +1 refers to the positive sentiment. As a result, each comment was submitted to TextBlob's sentiment analysis function through a Python script, resulting in an average polarity rating of comments made on pull requests by the LGBTQIA+ community on GitHub.

[2]https://textblob.readthedocs.io/en/dev/

4.10 Collecting Users Outside the Community

The method used for data collection followed the same approach as employed in the collection of users from the LGBTQIA+ community, totaling 4167 users, the same number as the users collected from the LGBTQIA+ community. To ensure equity in the analysis, the collection of users from outside the community was limited to this same number. A script developed in Python was utilized, integrating libraries such as 'requests' to make requests to the GitHub API. This script was responsible for performing GraphQL queries on the GitHub API, collecting data on issue comments and pull request comments from specific users for the analysis of the research questions (RQs), and storing them in a structured manner in MongoDB, ensuring the integrity and consistency of the collected data.

5 Results

In this section, the results obtained for each of the research questions are presented. To this end, 4.161 user profiles and 57.769 repositories were collected. In addition, 164.4 thousand pull request comments were collected. First of all, the data related to the users' technical profile is described. Next, the data from the repositories and the social aspect of the LGBTQIA+ community on GitHub. Finally, the data related to the analysis of feelings and the comparisons between the technical profile of community members and those who are not part of it.

5.1 What is the technical profile of GitHub users belonging to the LGBTQIA+ community?

5.1.1 Most used languages

The analysis of Figure 2 shows the ten most popular programming languages among these users. JavaScript is the most widely used programming language with 765 users, followed by Shell with 648 users, and Python comes in third with 565 users.

5.1.2 Frequency of commits

As shown in Figure 3, it is possible to observe a concentration of the data sample close to the median of 0 commits per day and week. It is also possible to see outliers of 10.45 commits per day and 73.14 commits per week.

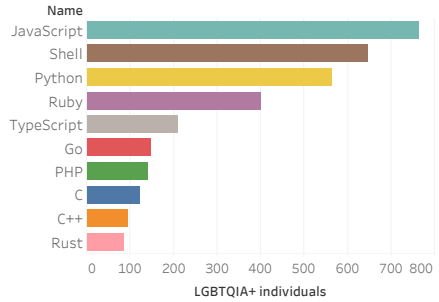

Figure 2. Number of usages of programming languages by LGBTQIA+ individuals

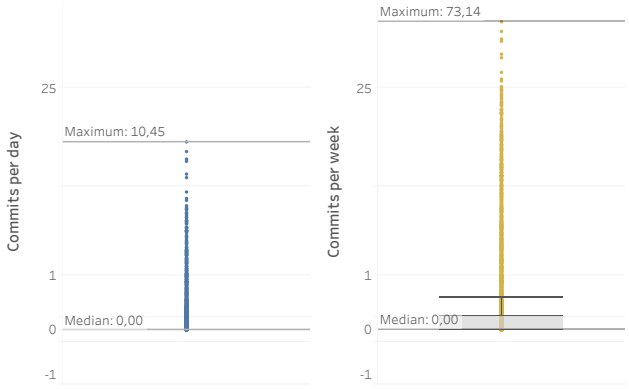

Figure 3. Number of commits per day and week, per user

5.1.3 Number of issues and pull requests

In the set of issues and pull requests, represented by the graphs in Figure 4, the median value of 0 is observed for both. Regarding the issues, the median and the lower quartile have a value of 0, while the upper quartile is equal to 2. This indicates that the majority of users, that is, around 75% of them, have less than 2 issues created. Additionally, outlier values of 3.584 issues and 7.710 pull requests were removed from the graph representation to prevent imprecise or incorrect conclusions about the dataset.

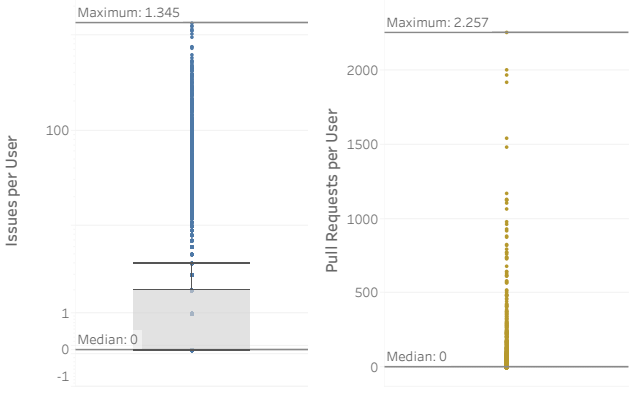

Figure 4. Number of issues and pull requests per user

5.2 What are the characteristics of repositories to which GitHub users belonging to the LGBTQIA+ community contribute?

5.2.1 Number of stargazers

Amongst the approximately 52 thousand repositories collected, approximately 41.04 thousand do not have any stars, which corresponds to a percentage of approximately 79% of all repositories. However, within this set, there are eight repositories in which users of the LGBTQIA+ community participate that have more than 10 thousand stars, as shown in Table 1.

Table 1. Number of stargazers per repository

Repository Name	Number of Stargazes
LAION-AI/Open-Assistant	32358
SerenityOS/serenity	27049
nushell/nushell	24560
syl20bnr/spacemacs	22963
llvm/llvm-project	19636
nrwl/nx	17495
sveltejs/kit	14486
ManimCommunity/manim	14268
ManimCommunity/manim	14267
rust-lang/book	12194
cfug/dio	12077

5.2.2 Primary language

The analysis of the 52.31 thousand repositories revealed the ten languages most widely used by users of the LGBTQIA+ community. As illustrated in Figure 5, JavaScript is the most predominant language, present in more than 10 thousand repositories. Python appears in more than 6 thousand repositories, and Ruby is present in about 3.055 of them.

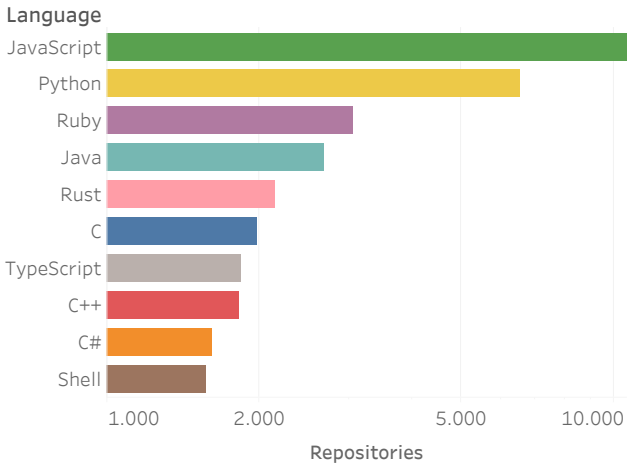

Figure 5. Number of repositories by language

5.2.3 Closed issues/Total issues

The density of closed issues per repository is calculated by dividing the total number of closed issues (173.36 thousand) by

the total number of issues (235.10 thousand). With the calculations, the result is 0.73, which indicates that the proportion of closed issues is 73%.

5.3 What is the social profile of GitHub users belonging to the LGBTQIA+ community amongst themselves?

5.3.1 Number of followers and followed people from the community

Figure 6 shows the distribution of the number of followers and followed profiles collected employing a box plot with a logarithmic scale. The graph reveals that 25% of the profiles have no followers, while 50% of them have no followers or do not follow other profiles. In addition, 75% of the analyzed users have up to 4 followers or follow up to 4 people.

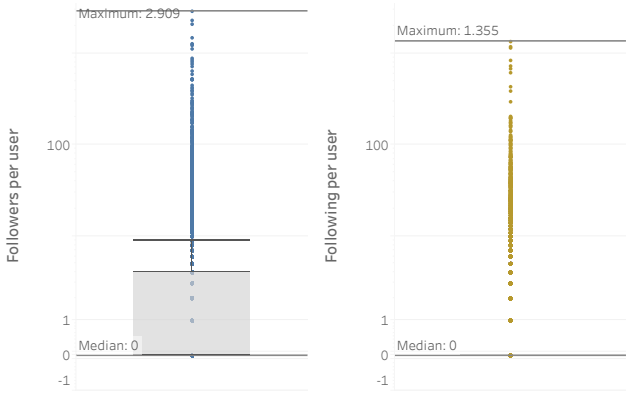

Figure 6. Number of followers and profiles followed by user

5.3.2 Number of sponsors and sponsored people from the community

Figure 7 shows the distribution of the number of sponsors and sponsored users per profile. It is observed that the lower quartile, the median, and the upper quartile are at 0, indicating that at least 75% of the sampled profiles do not have sponsors nor do they sponsor other users.

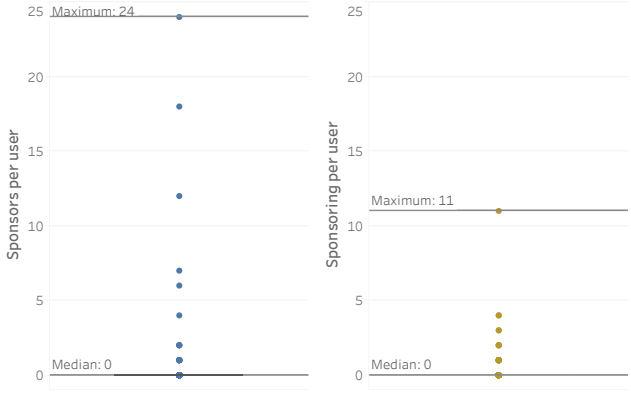

Figure 7. Number of sponsors and profiles sponsored by user

5.3.3 Areas with the most users from the LGBTQIA+ community

Figure 8 shows the top ten countries with valid profiles in the sample. It is important to note that the location field on GitHub allows any text, valid or not. A total of 1,787 profiles had invalid information in this field, therefore, the data shown in the figure is based on valid fillings.

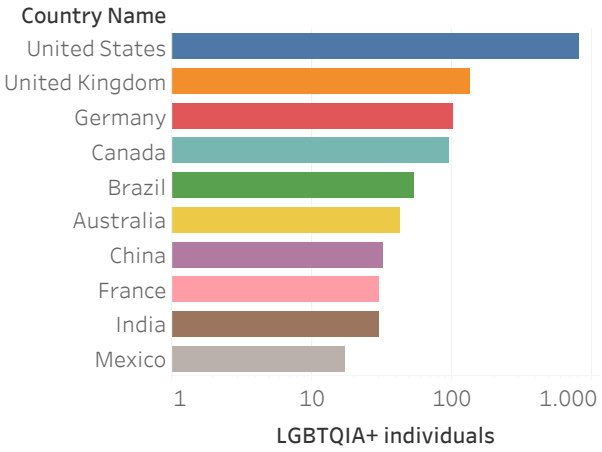

Figure 8. Number of LGBTQIA+ individuals by country

5.4 What are the sentiments related to comments on pull requests made by the LGBTQIA+ population on GitHub?

5.4.1 Average percentage of positive comments

Figure 9 displays the word cloud with positive comments extracted from the 164.4 thousand pull request comments. In it, it is possible to observe the most frequent words, highlighting “lib” (48 times), “shell” (47 times), and “python” (46 times). Besides, words like “work,” “thank,” and “good” suggest positive reviews. The average of these positive comments is 45.14%.

Figure 9. WordCloud of Positive Reviews

5.4.2 Average percentage of negative comments

By analyzing the comments in Figure 10, it is possible to identify that the words most frequently associated with negative aspects are “alpine” (27 times), “passwords” (26 times), “hard coded” (25 times), and “windows-server core” (24 times). The average of these negative comments is at a 15.88% percentage.

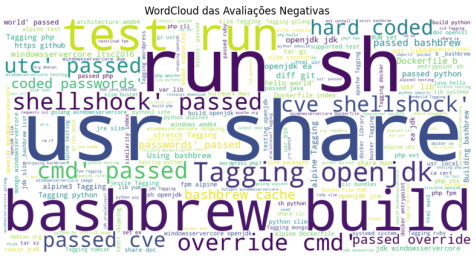

Figure 10. WordCloud of Negative Reviews

5.4.3 Average percentage of neutral comments

As indicated in Figure 11 with the neutral comments, words such as “ok”, “fine” and “average” suggest a neutral tone. Among the most frequent words, “Dockerfile” (20 times), “php” (19 times), and “linux” (18 times) stand out. The percentage of neutral comments observed is 38.97%.

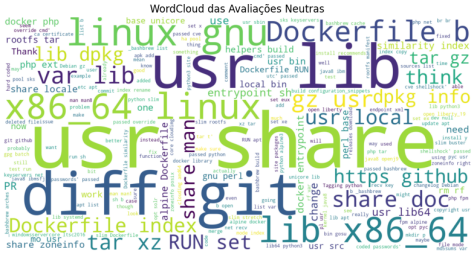

Figure 11. WordCloud of Neutral Reviews

5.5 What is the difference in the technical profile of people belonging to the LGBTQIA+ community and those who do not?

5.5.1 Most used languages

The detailed analysis of Figure 12 provides insight into the top ten programming languages utilized by users outside the community. Notably, Shell claims the top spot, showcasing its widespread usage, followed closely by JavaScript in second place, and Python securing the third position.

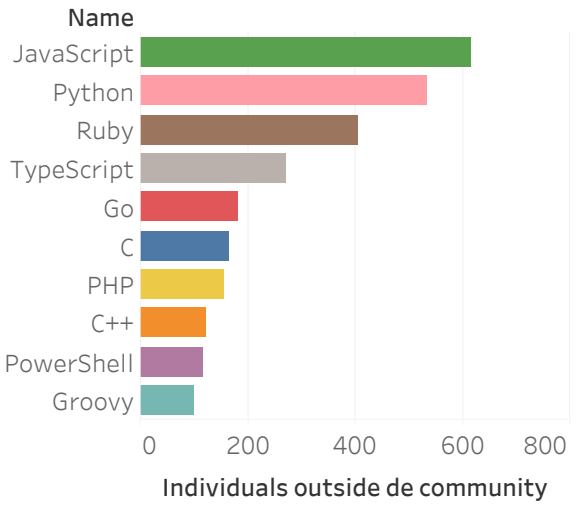

Figure 12. Number of programming languages usage by individuals outside the community

5.5.2 Closed issues/Total issues of the used repositories

Based on the Figure 13, it is observed that the majority of open issues, 80% of them, have been closed, and the total number of issues remains relatively stable over time.

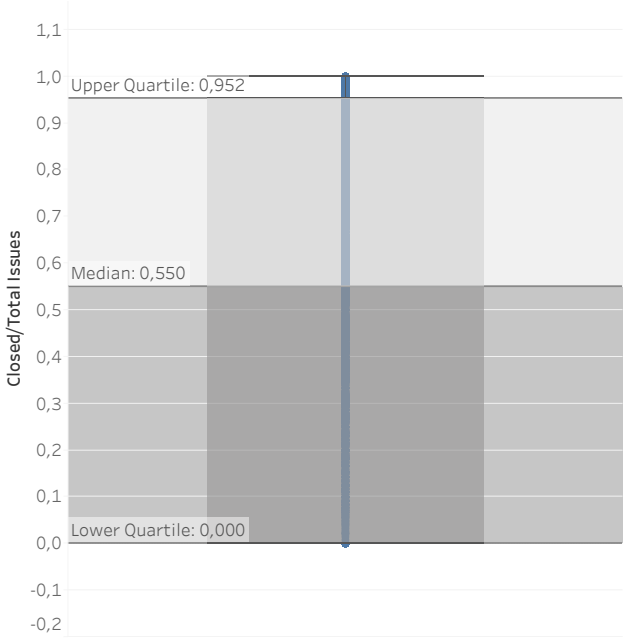

Figure 13. Ratio of closed issues to the total number of issues in repositories used by users outside the community

5.5.3 Number of issues and pull requests

In the Figure 14, it is possible to observe a relatively low number of issues and pull requests, while a few users exhibit a significantly higher quantity. This disparity is more pronounced for pull requests than for issues.

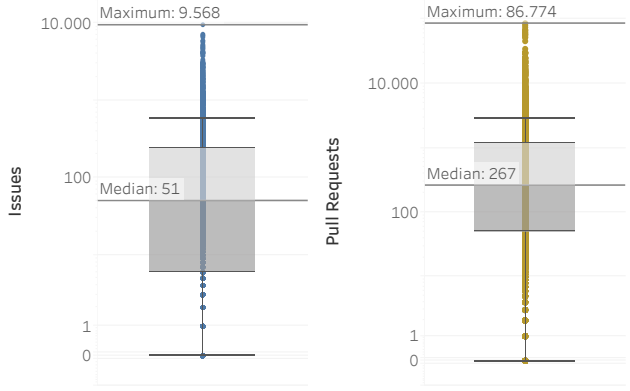

Figure 14. Number of issues and pull requests by users outside the community

6 Discussion

In this section, the results obtained for each research question are discussed. The study consisted of collecting GitHub users who are part of the LGBTQIA+ community to characterize the behavior of these profiles.

RQ.1 - What is the technical profile of GitHub users belonging to the LGBTQIA+ community?

With the collection of 4,161 profiles, it is evident that the users collected differ from the general GitHub audience in terms of the most used programming languages. Notably, Java, which is the third most used language on GitHub, is not among the top ten in the collected sample (GitHub Octoverse, 2022). Additionally, the analysis of data from commits, pull requests, and issues reveals that a significant portion of these profiles is inactive or uses GitHub primarily as an information source. At least 75% of the profiles have not made commits, pull requests, or issues for a year. This data can be particularly useful for companies looking to enhance diversity or create targeted inclusion programs. For instance, preparatory courses focused on languages that are less commonly used by specific groups could be developed to foster greater inclusivity and representation.

RQ.2 - What are the characteristics of repositories to which GitHub users belonging to the LGBTQIA+ community contribute?

First of all, concerning the repositories of these users, it is observed that approximately 41 thousand repositories, among the 57,769 total repositories, have zero stars, indicating that at least 78% of the repositories are not popular. In addition, it is interesting to point out that the most frequent primary language in this set is JavaScript, followed by Python. As a result, users from the LGBTQIA+ community contribute to repositories with the two most used primary languages on GitHub, according to data collected by GitHub (GitHub Octoverse, 2022).

Finally, it is important to note that repositories to which LGBTQIA+ community users contribute have, for the most part, little relevance in terms of popularity. However, there are notable exceptions, such as large repositories where there has been significant interaction. One example is LAION-AI/Open-Assistant, a project that provides access to a chat-based broad language model. This highlights that while many repositories may not receive much attention, there are still impactful projects involving contributions from LGBTQIA+ users.

RQ.3 - What is the social profile of GitHub users belonging to the LGBTQIA+ community amongst themselves?

When analyzing the data collection, it is noted that users from the LGBTQIA+ community do not frequently follow each other. However, some users stand out for having a high number of followers, even if they are not necessarily members of the community. Specifically, 0.57% of profiles have at least 500 followers, indicating a significant following for a small subset of users. When it comes to sponsorships, there are no meaningful interactions or financial support exchanges between members of the LGBTQIA+ community either.

As for the location of users, there is great diversity around the world due to the open text field used to collect the data, showing that LGBTQIA+ GitHub users are globally dispersed. This highlights an opportunity for organizations that support the LGBTQIA+ community to promote actions that increase the integration and development of these individuals within the tech environment. For example, creating a dedicated repository with courses and a Discord channel for these profiles to interact could facilitate the prospecting of contributors and increase inclusion in companies. This environment would provide a space to share job selection processes, discover talents, and give visibility to projects that need sponsorship. This approach can help foster a more supportive and interconnected community, enhancing opportunities for collaboration and career growth.

RQ.4 - What are the sentiments related to comments on pull requests made by the LGBTQIA+ population on GitHub?

When analyzing the 164.4k comments on pull requests from members of the LGBTQIA+ community, it is possible to observe that 45.14% of these messages reflect positive feelings. This ratio suggests that developers are frequently sharing constructive feedback optimistically, with no negative elements, which indicates a collaborative and encouraging dynamic within the community. This results includes, for example, praise for a job well done, constructive feedbacks and thanks for the helpful support provided. On the other hand, 15.88% of the comments expressed negative feelings. This may suggest expressions of dissatisfaction or concern about certain aspects of pull requests, or the interventions of other developers. This data highlights aspects that may need attention or improvement to strengthen the collaborative experience among community members, and this work can include actions by individual community members, as well as collective actions by social entities, or even inclusive policies that encourage professional growth developed by organizations involved in software construction and maintenance. The percentage of neutral sentiments, accounting for 38.97% of comments, indicates that many developers share information or perform tasks without significant emotion. This category can encompass the objective communication of data, updates, or the performance of routine tasks.

Considering these numbers, it is possible to infer that the dynamics of interaction between LGBTQIA+ developers on GitHub, in its essence, are predominantly positive, as most of the comments are loaded with constructive feelings, a promising and optimistic view of the scenario of the participation of these developers in the general context. However, the presence of negative feedback suggests areas of opportunity for improvement and greater alignment in collaboration among organization members. This insight can be valuable for fostering a more supportive and cohesive environment for LGBTQIA+ developers on the platform.

RQ.5 - What is the difference in the technical profile of people belonging to the LGBTQIA+ community and those who do not?

When comparing the technical profiles of community members and non-members, it is evident that JavaScript is the predominant language in both groups. However, notable divergences arise in technological preferences. For instance, the frequent use of Groovy by non-members is a significant difference, as Groovy is absent from the top ten languages used by LGBTQIA+ members. Similarly, Shell is widely utilized by community members but does not stand out among non-members, reflecting a diversity in technological preferences between the two groups.

The high issue resolution rate is a positive indicator of ef-

fectiveness in both groups. Specifically, 73% of issues were closed by community members, while 80% were closed by non-members. This analysis does not reveal a significant difference, indicating similar effectiveness in issue resolution between both groups.

In addition to this analysis, it is noted that non-community members also exhibit low activity regarding issues and pull requests, a pattern similar to that identified among community members. This similarity suggests that activity levels in managing issues and pull requests are low across the board, regardless of community affiliation. This insight can be valuable for understanding the broader patterns of engagement and contribution on GitHub, and for identifying areas where increased activity and participation could be encouraged.

To illustrate, the maintainers of the Groovy language mentioned above could encourage projects and challenges aimed at the community as a way to increase the participation of these developers in scenarios where this technology is used. Or, projects using JavaScript could take advantage of its unquestionable popularity to create campaigns and/or events, in order to provide space for community members to serve as examples for those who do not see themselves as belonging to that environment.

7 Threats to Validity

In this section, the threats to the validity of this study are presented, as well as the strategies adopted to mitigate them.

First of all, as for the validity of its construction, the size of the total LGBTQIA+ community population registered on GitHub is unknown. To mitigate this threat, the 19 keywords sought to include as many terms used for the self-identification of LGBTQIA+ people as possible.

Regarding internal validity, the existence of keywords in GitHub user profiles does not guarantee that one is using them as a self-identification term. This can lead to the collection of users who do not belong to the LGBTQIA+ community. To confront this threat, a script was created to filter users based on their usage of the keywords.

As for external validity, the generalization of collected data is not feasible, as is the case with many other Software Engineering studies. This is due to the possibility that the data does not represent the diverse contexts of all minority groups inside the LGBTQIA+ community. However, this threat is mitigated with this study's comprehensive analysis, which comprehended 57.769 repositories and 4.161 users collected through 19 keywords.

Finally, regarding the threat to conclusion validity, it is worth noting the limitations of the characterization carried out in relation to irregularities in the identification and classification of the people collected. Therefore, it is possible that the characterization concluded from the analysis of the profiles is not entirely assertive. This threat is mitigated by taking into account various properties of the technical, social and interactive profiles of the people collected.

8 Conclusion

In this work, a study was carried out in order to characterize people from the LGBTQIA+ community registered on the GitHub platform. When analyzing the profiles obtained according to the technical settings, JavaScript is observed as the most widely used language. Looking at the repositories to which the community contributes, it is noted that 78% of them do not have stars. The social aspect was also analyzed, and no evidence of interaction between profiles from the community was found. Among the collected users, there are no followers and followed people among the collected sample. When looking at sponsorships and sponsors, the behavior is the same.

The feelings expressed in the comments reveal a predominantly positive dynamic in the interaction between users. However, even though there is a low percentage of negative comments, it is possible to identify space for improvement, thus contributing to an even healthier environment.

When analyzing community members and non-members, it becomes apparent that disparities exist in preferences for technological languages. Furthermore, the participation of non-members in repositories proves to be an area requiring attention for enhancement. In this context, there is an identified need to improve existing dynamics and foster a more extensive and diversified collaboration between these two participant groups.

Therefore, organizations that seek to promote diversity and inclusion of the LGBTQIA+ community must devise strategies for the integration of these people. The data presented provides a technical and social direction, such as which languages are popular among this minority and languages whose learning can still be stimulated, with room for growth.

For future studies, it would be interesting to further analyze keywords to identify new individuals from the LGBTQIA+ community. Methods such as interviews or surveys could mitigate internal validity threats. Investigating scenarios of possible discrimination against LGBTQIA+ users that were not addressed by the platform is also a prospect. Finally, additional insights for RQ5, including metrics such as the 'number of commits', will be explored to provide a more comprehensive comparison of the technical profiles between the LGBTQIA+ community and non-members.

References

Aberdour, M. (2007). Achieving quality in open-source software. *IEEE Software*, 24(1):58–64.

Albusays, K., Bjorn, P., Dabbish, L., Ford, D., Murphy-Hill, E., Serebrenik, A., and Storey, M.-A. (2021). The diversity crisis in software development. *IEEE Software*, 38(2):19–25.

Almarzouq, M., Alzaidan, A., and AlDallal, J. (2020). Mining github for research and education: challenges and opportunities. *International Journal of Web Information Systems*, ahead-of-print.

Aué, J., Haisma, M., Tómasdóttir, K. F., and Bacchelli, A.

(2016). Social diversity and growth levels of open source software projects on github. In *Proceedings of the 10th ACM/IEEE International symposium on empirical software engineering and measurement*, pages 1–6.

Bosu, A. and Sultana, K. Z. (2019). Diversity and inclusion in open source software (oss) projects: Where do we stand? In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–11.

Casalnuovo, C., Vasilescu, B., Devanbu, P., and Filkov, V. (2015). Developer onboarding in github: the role of prior social links and language experience. In *Proceedings of the 2015 10th joint meeting on foundations of software engineering*, pages 817–828.

DuBow, W. M., Jones, S., Sexton, S., and Tadimalla, S. Y. (2024). Broadening participation in computing education: Advancing lgbtqia+ voices. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSE 2024, page 1529–1530, New York, NY, USA. Association for Computing Machinery.

Dutta, R., Costa, D. E., Shihab, E., and Tajmel, T. (2023). Diversity awareness in software engineering participant research. *SEIS - Software Engineering in Society*, ahead-of-print.

Eaton, M. E. (2018). A comparative analysis of the use of github by librarians and non-librarians. *Evidence Based Library and Information Practice*, 13(2):27–48.

Etaiwi, W. and Naymat, G. (2017). The impact of applying different preprocessing steps on review spam detection. *Procedia Computer Science*, 113:273–279. The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops.

Garcia, R., Treude, C., and La, W. (2023). Towards understanding the open source interest in gender-related github projects. *arXiv preprint arXiv:2303.09727*.

Gila, A. R., Jaafa, J., Omar, M., and Tunio, M. Z. (2014). Impact of personality and gender diversity on software development teams' performance. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, pages 261–265.

GitHub Octoverse (2022). Octoverse: Top programming languages. Website. Acesso em: 24 de maio de 2023.

Guzman, E., Azócar, D., and Li, Y. (2014). Sentiment analysis of commit comments in github: an empirical study. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, page 352–355, New York, NY, USA. Association for Computing Machinery.

Haddi, E., Liu, X., and Shi, Y. (2013). The role of text preprocessing in sentiment analysis. *Procedia Computer Science*, 17:26–32. First International Conference on Information Technology and Quantitative Management.

Hu, Y., Wang, S., Ren, Y., and Choo, K.-K. R. (2018). User influence analysis for github developer social networks. *Expert Systems with Applications*, 108:108–118.

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

Janzen, D. S., Bahrami, S., Silva, B. C. d., and Falessi, D. (2018). A reflection on diversity and inclusivity efforts in a software engineering program. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–9.

Kumar, A., Khare, M., and Tiwari, S. (2022). Sentiment analysis of developers' comments on github repository: A study. In *2022 14th International Conference on Advanced Computational Intelligence (ICACI)*, pages 91–98.

Lee, E., Rustam, F., Washington, P. B., Barakaz, F. E., Aljedaani, W., and Ashraf, I. (2022). Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. *IEEE Access*, 10:9717–9728.

Montandon, J. E., Valente, M. T., and Silva, L. L. (2021). Mining the technical roles of github users. *Information and Software Technology*, 131:106485.

Paiva, E., Carvalho, G., Mayrink, J., Maruch, M., Felix, P., Pacheco, G., and Xavier, L. (2023). Caracterização da população lgbtqia+ na plataforma github. In *Anais do XI Workshop de Visualização, Evolução e Manutenção de Software*, pages 16–20, Porto Alegre, RS, Brasil. SBC.

Rehman, I., Wang, D., Kula, R. G., Ishio, T., and Matsumoto, K. (2020). Newcomer candidate: Characterizing contributions of a novice developer to github.

Richard, T. S., Wiese, E. S., and Rakamarić, Z. (2022). An lgbtq-inclusive problem set in discrete mathematics. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1*, pages 682–688.

Saxena, R. and Pedanekar, N. (2017). I know what you coded last summer: Mining candidate expertise from github repositories. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 Companion, page 299–302, New York, NY, USA. Association for Computing Machinery.

Shimada, N., Xiao, T., Hata, H., Treude, C., and Matsumoto, K. (2022). Github sponsors: exploring a new way to contribute to open source. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, page 1058–1069, New York, NY, USA. Association for Computing Machinery.

Sultana, S., Uddin, G., and Bosu, A. (2024). Assessing the influence of toxic and gender discriminatory communication on perceptible diversity in oss projects.

van der Meulen, M. J. and Revilla, M. A. (2008). The effectiveness of software diversity in a large population of programs. *IEEE Transactions on Software Engineering*, 34(6):753–764.

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

Vasilescu, B., Posnett, D., Ray, B., van den Brand, M. G., Serebrenik, A., Devanbu, P., and Filkov, V. (2015b). Gender and tenure diversity in github teams. In *Proceedings*

of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, page 3789–3798, New York, NY, USA. Association for Computing Machinery.

Wall, S. (2023). *The development of the LGBT+ community in the UK in the last 50 years*, page 17–27. Routledge.

Wang, J. and Hejazi Moghadam, S. (2017). Diversity barriers in k-12 computer science education: Structural and social. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*, pages 615–620.

Wang, Y., Wang, L., Hu, H., Jiang, J., Kuang, H., and Tao, X. (2022). The influence of sponsorship on open-source software developers' activities on github. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 924–933.

Zhao, S., Xia, X., Fitzgerald, B., Li, X., Lenarduzzi, V., Taibi, D., Wang, R., Wang, W., and Tian, C. (2024). Openrank leaderboard: Motivating open source collaborations through social network evaluation in alibaba. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, ICSE-SEIP '24, page 346–357, New York, NY, USA. Association for Computing Machinery.