Caracterização da população LGBTQIA+ na plataforma GitHub

Erick Paiva
Depto. de Engenharia de Software
PUC Minas
Belo Horizonte, Brasil
erick.paiva@sga.pucminas.br

Guilherme Carvalho
Depto. de Engenharia de Software
PUC Minas
Belo Horizonte, Brasil
blcpinto@sga.pucminas.br

João Pedro Mayrink
Depto. de Engenharia de Software
PUC Minas
Belo Horizonte, Brasil
joao.jesus.1302911@sga.pucminas.br

Maria Clara Maruch
Depto. de Engenharia de Software
PUC Minas
Belo Horizonte, Brasil
maria.jabali@sga.pucminas.br

Pedro Felix
Depto. de Engenharia de Software
PUC Minas
Belo Horizonte, Brasil
pedro.costa.1298438@sga.pucminas.br

Gabriel Pacheco
Depto. de Engenharia de Software
PUC Minas
Belo Horizonte, Brasil
gacampacheco@gmail.com

Laerte Xavier
Depto. de Engenharia de Software
PUC Minas
Belo Horizonte, Brasil
laertexavier@pucminas.br

RESUMO

Representatividade e inclusão são preocupações atuais da Engenharia de Software. Para lidar com esse assunto, estudos anteriores apresentam dados e discutem estratégias para mitigar a desigualdade na formação de times. Entretanto, em projetos de código-aberto, ainda não está claro como essas preocupações permitem a inclusão de pessoas desenvolvedoras pertencentes a grupos de minoria. Particularmente, pouco se sabe a respeito da participação de pessoas da comunidade LGBTQIA+ no desenvolvimento de projetos populares do GitHub. Diante disso, este estudo visa caracterizar a participação dessa comunidade, a partir da coleta de 3K perfis de usuários e 52K repositórios. Busca-se entender o comportamento e a interação desses usuários dentro desse contexto, assim como provocar a criação de estratégias que promovam maior inclusão. Como resultado, foi observado que o perfil técnico dos usuários apresenta afastamento da comunidade desenvolvedora geral, apesar da sua participação ativa em repositórios populares de JavaScript e Python. O perfil social é isolado, com poucas interações entre os usuários, embora haja concentração em áreas específicas no mapa da plataforma.

KEYWORDS

diversidade, lgbtqia+, mineração de repositórios, perfis, interação

1 INTRODUÇÃO

Na graduação em engenharia de software, indivíduos que são lésbicas, gays, bissexuais, transgêneros, *queer*, intersexuais, assexuais, entre outras identificações de sexualidade e gênero (LGBTQIA+) estão entre os grupos sub-representados que frequentemente enfrentam sentimentos de exclusão e marginalização [6]. Mesmo sendo um cenário complexo que envolve aspectos históricos, sociais e culturais, é importante salientar que ao longo dos anos foram propostos estudos e estratégias para enfrentar a desigualdade [4, 8].

O GitHub[1] é uma plataforma que possibilita aos desenvolvedores compartilhar projetos de software [4]. Com o tempo, a plataforma se estabeleceu como referência para estudos abrangentes sobre organização técnica e social [7]. Porém, pouco se sabe sobre a participação da comunidade LGBTQIA+ no GitHub. Portanto, o problema abordado nesta pesquisa é a escassez de estudos específicos que analisem e compreendam a contribuição dessas pessoas na plataforma GitHub.

Nesse contexto, a motivação para realizar esta pesquisa é obter um conhecimento mais aprofundado sobre a comunidade LGBTQIA+ no GitHub, visto que a literatura atual carece de estudos abrangentes dessa área. Assim, este estudo se justifica pelo fato de que organizações e proprietários de repositórios podem utilizar as informações obtidas para tomar decisões estratégicas sobre como atrair mais membros dessa comunidade para seus projetos.

Assim, o objetivo geral deste trabalho é caracterizar a participação da população LGBTQIA+ na plataforma GitHub, apresentando um panorama sobre essa população no GitHub. Para atingir esse objetivo, são propostos as seguintes *research questions* (RQs):

- RQ.1 Qual o perfil técnico das pessoas do GitHub pertencentes à comunidade LGBTQIA+?
- RQ.2 Quais as características dos repositórios em que pessoas do GitHub pertencentes à comunidade LGBTQIA+ contribuem?
- RQ.3 Qual o perfil social das pessoas do GitHub pertencentes à comunidade LGBTQIA+ entre si?

O restante deste trabalho está organizado da seguinte maneira: a Seção 2 descreve a metodologia utilizada para o estudo. Na Seção 3, apresentam-se os resultados obtidos. A Seção 4 discute as questões de pesquisa propostas. A Seção 5 detalha as ameaças à validade e suas mitigações. Por fim, as Seções 6 e 7 apresentam os trabalhos relacionados e a conclusão deste estudo, respectivamente.

[1] https://github.com

2 METODOLOGIA

O trabalho proposto é classificado como uma pesquisa de caracterização, cujo objetivo é descrever as características de uma população ou fenômeno. Para cada RQ, foram definidas as seguintes métricas:

- RQ.1 Qual o perfil técnico das pessoas do GitHub pertencentes à comunidade LGBTQIA+?
  - M.1 Linguagens mais utilizadas;
  - M.2 Frequência de *commits*;
  - M.3 Quantidades de *issues* e *pull requests*;
- RQ.2 Quais as características dos repositórios em que pessoas do GitHub pertencentes à comunidade LGBTQIA+ contribuem?
  - M.1 Quantidade de estrelas;
  - M.2 Linguagem primária;
  - M.3 *Issues* fechadas/total de *issues*;
- RQ.3 Qual o perfil social das pessoas do GitHub pertencentes à comunidade LGBTQIA+ entre si?
  - M.1 Número de seguidores e pessoas seguidas da comunidade;
  - M.2 Número de patrocinadores e pessoas patrocinadas da comunidade;
  - M.3 Regiões com mais usuários da comunidade;

No restante desta seção, são descritos os procedimentos (Seção 2.1), os métodos adotados para mineração dos usuários do GitHub (Seção 2.2) e para a coleta dos repositórios (Seção 2.3).

2.1 Procedimentos

A Figura 1 apresenta uma visão geral do processo adotado neste trabalho, dividido em três etapas principais: coleta de dados de usuários LGBTQIA+ no GitHub, coleta de dados dos repositórios em que esses usuários contribuem e análise e cálculo das métricas. Para isso, foram criados *scripts* Python e os dados foram armazenados no banco de dados MongoDB.

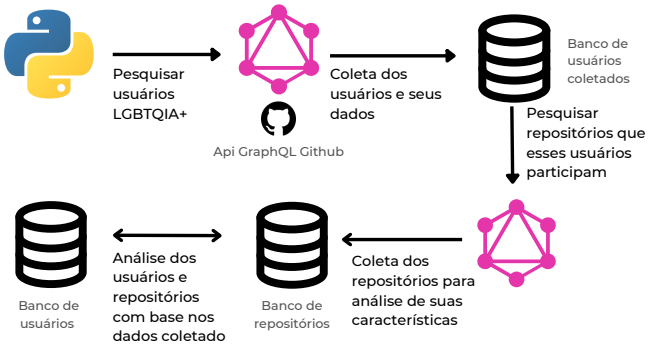

Figura 1: Visão geral da metodologia adotada

2.2 Mineração dos Usuários

Para a busca de usuários que se autoidentificam como pertencentes à comunidade LGBTQIA+ dentro da plataforma, foram utilizadas palavras-chave como o filtro para os campos de nome e biografia no perfil. Assim, foram selecionados os usuários que incluíam os seguintes termos: "queer", "rainbow_flag", "transgender_flag", "non-binary", "non binary", "lesbian", "bisexual", "asexual", "pansexual", "transgender", "they them", "he them", "she them", "gay", "trans", "transboy", "transgirl", "transwoman" e "transmen".

Para validar os usuários obtidos na primeira busca, foi desenvolvido um *script* em Python que filtrava os usuários, garantindo a presença das palavras-chave utilizadas. O *script* verificava se o usuário coletado possuía a palavra-chave no nome ou na biografia, de forma independente, sem estar inserida em outra palavra (i.e., não foram selecionados perfis que continham esses termos como *substring*). Além disso, realizou-se a verificação de duplicidade entre os usuários coletados, já que um usuário poderia ser retornado em várias consultas, se contivesse mais de uma palavra-chave em seu perfil. A remoção dessas duplicatas se deu por meio da análise do *login* (i.e., nome de usuário) de cada perfil retornado na consulta.

Durante a busca inicial, os dados de todos os usuários encontrados foram coletados para traçar seus perfis técnicos. Em seguida, foi desenvolvido um *script* em Python, utilizando a API GraphQL do GitHub, para obter os *commits*, *pull requests* e as linguagens de programação contidas nos *commits* de cada *pull request* realizado pelos usuários (métricas definidas para a RQ.1). Por fim, as demais métricas foram calculadas por meio da identificação dos repositórios nos quais esses usuários são ativos (métricas para a RQ.2), e da análise da lista de seguidores de cada usuário (métricas para a RQ.3).

2.3 Coleta dos Repositórios

Para avaliar as características dos repositórios em que usuários LGBTQIA+ contribuem, foi coletado os repositórios com base nos perfis identificados na mineração inicial. Isso envolve outra consulta usando a API GraphQL do GitHub para obter esses repositórios em que eles contribuíram. Utiliza-se os *login* dos usuários da primeira coleta como identificação para essa coleta. Nesse processo, são extraídas informações como idade do repositório, quantidade de estrelas, *issues*, *commits*, data do último *commit* e linguagem principal.

3 RESULTADOS

Nesta seção, são apresentados os resultados obtidos para cada uma das *research questions*. Para tanto, foram coletados 3.175 perfis de usuários e 52.516 repositórios. Primeiro, descrevem-se os dados relativos ao perfil técnico dos usuários. Em seguida, os dados dos repositórios e do aspecto social da comunidade LGBTQIA+ no GitHub.

3.1 Qual o perfil técnico das pessoas do GitHub pertencentes à comunidade LGBTQIA+?

3.1.1 Linguagens mais utilizadas. A análise da Figura 2 é visto as dez linguagens de programação mais populares entre esses usuários. JavaScript é a linguagem de programação mais utilizada, com 643 usuários, em seguida Shell, com 553 usuários e, em terceiro, Python, com 485 usuários.

3.1.2 Frequência de commits. Conforme a Figura 3, é possível observar uma concentração da amostra de dados próxima à mediana de 0 *commits* por dia e por semana. É possível ver, também, valores *outliers* de 10,45 *commits* por dia e 73,14 *commits* por semana.

3.1.3 Quantidades de issues e pull requests. No conjunto de *issues* e *pull requests*, representados pelos gráficos da Figura 4, observa-se

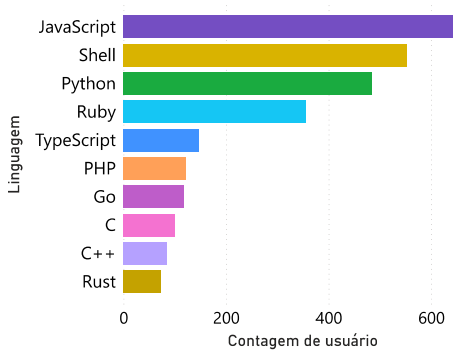

Figura 2: Número de utilização das linguagens de programação

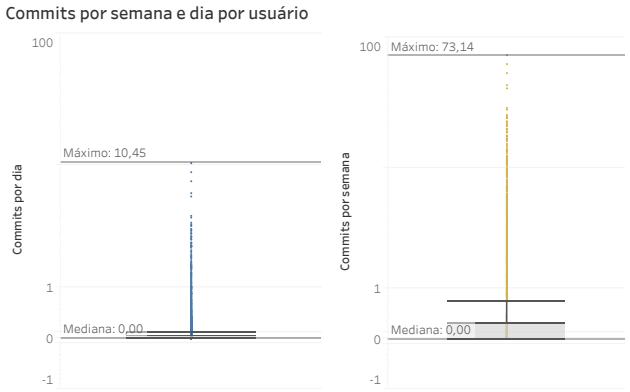

Figura 3: Quantidade de *commits* por dia e semana, por usuário

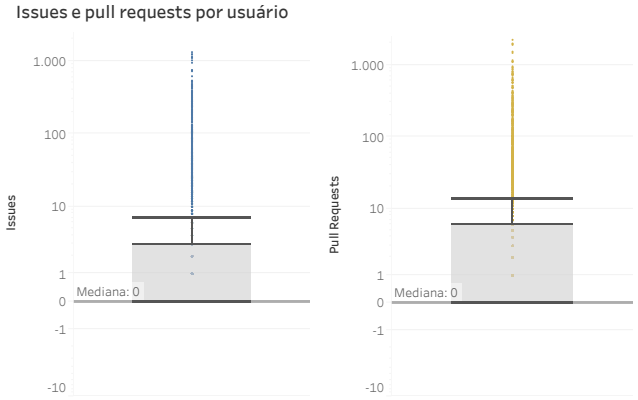

Figura 4: Quantidade de *issues* e *pull requests* por usuário

mediana com valor 0. Em relação às *issues*, a mediana e o primeiro quartil têm valor 0, enquanto o terceiro quartil é igual a 3. Isso indica que a maioria, cerca de 75%, tem menos de 3 *issues* criadas.

3.2 Quais as características dos repositórios em que pessoas do GitHub pertencentes à comunidade LGBTQIA+ contribuem?

Tabela 1: Número de estrelas por repositório

Nome do repositório	Quantidade de estrelas
LAION-AI/Open-Assistant	32358
ManimCommunity/manim	28535
nushell/nushell	24560
syl20bnr/spacemacs	22963
llvm/llvm-project	19636
nrwl/nx	17495
sveltejs/kit	14486
rust-lang/book	12194
rshipp/awesome-malware-analysis	9775
tianon/gosu	8722

3.2.1 Quantidade de estrelas. Dentre os cerca de 52 mil repositórios coletados, aproximadamente 41,04 mil repositórios não possuem nenhuma estrela, correspondendo a uma porcentagem de aproximadamente 79% do total de repositórios. No entanto, dentro desse conjunto, existem oito repositórios nos quais os usuários da comunidade LGBTQIA+ participam que possuem mais de 10 mil estrelas, como apresentado na Tabela 1.

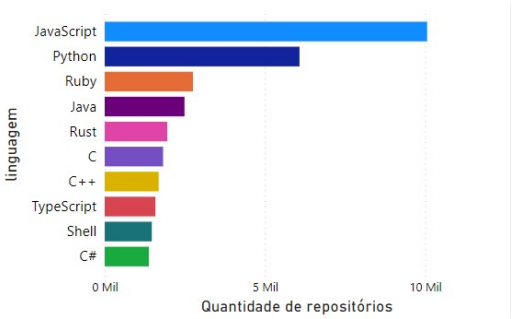

Figura 5: Contagem de repositórios por linguagem

3.2.2 Linguagem primária. A análise dos 52,31 mil repositórios revelou as dez linguagens mais utilizadas pelos usuários da comunidade LGBTQIA+. Como ilustrado na Figura 5, o JavaScript é a linguagem mais predominante, presente em mais de 10 mil repositórios. Python aparece em mais de 6 mil repositórios, e Ruby está presente em cerca de 2.800 repositórios.

3.2.3 Issues fechadas/total de issues. A densidade de *issues* fechadas por repositório é calculada dividindo o número total de *issues* fechadas (173,36 mil) pelo total de *issues* (235,10 mil). Com os cálculos, obtêm-se o resultado 0,73, o qual indica que a proporção de *issues* fechadas é de 73%.

3.3 Qual o perfil social das pessoas do GitHub pertencentes à comunidade LGBTQIA+ entre si?

3.3.1 Quantidade de seguidores e pessoas seguidas dentro da comunidade. A Figura 6 apresenta a distribuição da quantidade de

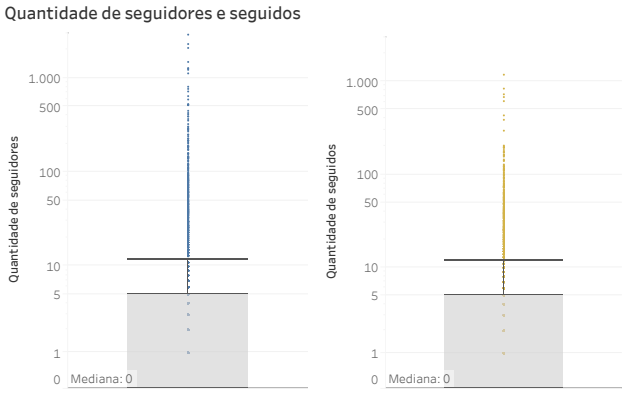

Figura 6: Número de seguidores e perfis seguidos por usuário

seguidores e seguidos dos perfis coletados por meio de um *box-plot* com escala logarítmica. O gráfico revela que 25% dos perfis não possuem seguidores, enquanto 50% não têm seguidores ou não seguem outros perfis. Além disso, 75% dos usuários analisados têm até 5 seguidores ou seguem até 5 pessoas.

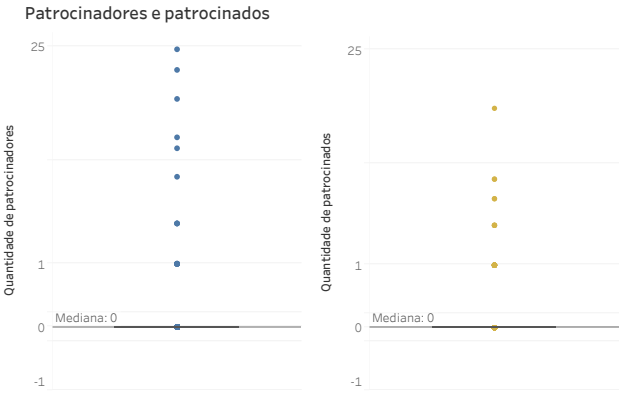

Figura 7: Número de patrocinadores e perfis patrocinados por usuário

3.3.2 Quantidade de patrocinadores e pessoas patrocinadas dentro da comunidade. A Figura 7 mostra a distribuição do número de patrocinadores e usuários patrocinados por perfil. Observa-se que o primeiro quartil, a mediana e o terceiro quartil estão em zero, indicando que pelo menos 75% dos perfis da amostra não possuem patrocinadores nem patrocinam outros usuários.

3.3.3 Regiões com mais perfis da comunidade LGBTQIA+. A Figura 8 apresenta os dez principais países com perfis válidos na amostra. É importante ressaltar que o campo de localização no GitHub permite qualquer texto, válido ou não. Foram identificados 1.787 perfis com preenchimentos inválidos nesse campo, portanto, os dados exibidos na figura são baseados nos preenchimentos válidos.

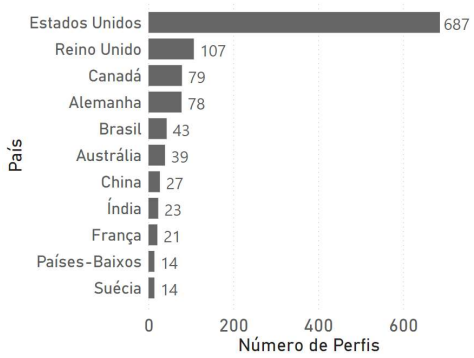

Figura 8: Número de perfis por país

4 DISCUSSÃO

Nesta seção, é apresentada a discussão dos resultados obtidos referentes a cada questão de pesquisa. O estudo realizado consistiu na coleta de pessoas usuárias no GitHub que fazem parte da comunidade LGBTQIA+ para a caracterização do comportamento desses perfis.

RQ1 - Qual o perfil técnico das pessoas do GitHub pertencentes à comunidade LGBTQIA+? Com a coleta de 3.157 perfis, as pessoas usuárias coletadas apresentam diferenças em relação ao público geral do GitHub em termos de linguagens de programação mais utilizadas. O Java, que é a terceira linguagem mais utilizada no GitHub, não está entre as dez primeiras na amostra coletada [5]. Além disso, a análise dos dados de *commits*, *pull requests* e *issues* revela que uma parte significativa dos perfis é inativa ou usa o GitHub apenas como fonte de consulta. Pelo menos 75% dos perfis não realizaram commits, pull requests ou issues no período de um ano. Esses dados podem ser úteis para empresas que buscam aumentar a diversidade ou criar programas de inclusão e diversidade, como cursos preparatórios voltados para linguagens menos utilizadas por grupos específicos.

RQ.2 Quais as características dos repositórios em que pessoas do GitHub pertencentes à comunidade LGBTQIA+ contribuem? Primeiramente, em relação aos repositórios destes usuários, observa-se que aproximadamente 41 mil repositórios, dentre os 52.310 repositórios totais, possuem zero estrelas, apontando que, no mínimo, 78% dos repositórios não são populares. Além disso, é interessante pontuar que a linguagem primária mais frequente desse conjunto é o JavaScript, seguido pelo Python. Com isso, usuários da comunidade LGBTQIA+ contribuem em repositórios com as duas linguagens primárias mais utilizadas do GitHub, conforme dados levantados pelo GitHub em 2022 [5]. Por fim, repositórios em que os usuários da comunidade contribuem possuem, em grande maioria, pouca relevância. Entretanto, há grandes repositórios em que houve alguma interação, como o `LAION-AI/Open-Assistant`, projeto que fornece acesso a um modelo de linguagem ampla baseado em chat.

RQ3 - Qual o perfil social das pessoas do GitHub pertencentes à comunidade LGBTQIA+ entre si? Ao analisar a coleta de dados, nota-se que os usuários da comunidade LGBTQIA+ não se seguem entre si. No entanto, alguns usuários se destacam por terem um alto número de seguidores, mesmo que não sejam necessariamente membros da

comunidade, 0,57% dos perfis possuem pelo menos 500 seguidores. No que diz respeito aos patrocínios, também não há interações significativas entre os membros da comunidade LGBTQIA+. Quanto à localização dos usuários, há uma grande diversidade em todo o mundo devido à natureza do campo de texto aberto utilizado para coletar os dados. Dessa forma, organizações apoiadoras da comunidade LGBTQIA+ podem produzir ações para aumentar a integração dessas pessoas e desenvolvê-las dentro do ambiente criado. Um repositório com cursos e um *discord* para interação desses perfis, para facilitar a prospecção de contribuidores e aumentar a inclusão em empresas, já que existiria um ambiente para divulgar processos seletivos e buscar talentos. Além disso, esse ambiente dará visibilidade a projetos que necessitam de patrocínio.

5 AMEAÇAS À VALIDADE

Nesta seção, são apresentadas as ameaças à validade deste estudo, assim como as estratégias adotadas para mitigá-las.

Primeiramente, quanto à validade de construção, o tamanho da população total da comunidade LGBTQIA+ cadastrada no GitHub não é conhecido. Para mitigar essa ameaça, as 19 palavras-chave buscaram incluir o maior número possível de termos utilizados para a autoidentificação de pessoas LGBTQIA+.

Em relação à validade interna, a existência de palavras-chave nos perfis dos usuários do GitHub não garante que esteja utilizando como termo de autoidentificação. Isso pode levar à coleta de pessoas não pertencentes à comunidade LGBTQIA+. Para enfrentar essa ameaça, foi criado um *script* que filtra os usuários com base no uso das palavras-chave.

Quanto à validade externa, a generalização dos dados coletados não é viável, assim como em muitos estudos na Engenharia de Software. Isso se deve à possibilidade de os dados não representarem os diversos contextos de todos os grupos minoritários na comunidade LGBTQIA+. No entanto, essa ameaça é mitigada devido à análise abrangente deste estudo, envolvendo 52.516 repositórios e 3.175 usuários, obtidos por meio de 19 palavras-chave.

6 TRABALHOS RELACIONADOS

A princípio, perfis engajados em projetos *open-source* no GitHub geram conteúdos como código e comentários [1] que podem ser acessados via Interface de Programação de Aplicação (API, do inglês *Application Programming Interface*). Portanto, é comum o uso da plataforma em estudo de caracterização como este, visto que ela fornece uma gama de informações sobre projetos, usuários e organizações, essenciais para a condução da pesquisa.

Além disso, perfis influentes, populares e ativos no GitHub atraem seguidores para novos projetos [2]. As pessoas seguem novos perfis em busca de aprendizado, socialização e colaboração, demonstrando como esses usuários participam na plataforma GitHub.

Por fim, outro trabalho identifica a diversidade de grupos sociais encontrada nas pesquisas na área de Engenharia de Software [3]. É realizada a coleta de 79 trabalhos de pesquisa que contém 105 estudos envolvendo a comunidade, sendo identificado no total 12 categorias de diversidade, o que evidência a existência de conteúdo a ser estudado sobre grupos sociais específicos na plataforma GitHub.

Em suma, observa-se que existe uma escassez de trabalhos *quantitativos* que caracterizam a comunidade LGBTQIA+ no GitHub. Assim, este trabalho fornece dados e análises que buscam ampliar a compreensão das dinâmicas dessa população na plataforma.

7 CONCLUSÃO

Neste trabalho, foi realizado um estudo a fim de caracterizar as pessoas da comunidade LGBTQIA+ cadastradas na plataforma GitHub. Ao analisar os perfis obtidos de acordo com a parte técnica, observa-se o JavaScript como a linguagem mais utilizada. Observando os repositórios que a comunidade contribui, nota-se que 78% deles não possuem estrelas. O último aspecto analisado foi o social, no qual não foram encontrados indícios de interação entre os perfis da comunidade. Das pessoas coletadas, não existem seguidores e pessoas seguidas dentre a amostra coletada. Ao observar os patrocínios e patrocinadores o comportamento se repete.

Logo, organizações que buscam promover a diversidade e a inclusão da comunidade LGBTQIA+ devem traçar estratégias para integração dessas pessoas. Os dados apresentados fornecem um direcionamento técnico e social, como as linguagens populares entre essa minoria e linguagens que o aprendizado ainda pode ser estimulado, com espaço para crescimento.

Para trabalhos futuros, planeja-se analisar mais profundamente as palavras-chave para identificar novos públicos da comunidade LGBTQIA+. Isso será combinado com métodos como entrevistas ou pesquisas para mitigar a validade interna. Pretende-se, ainda, investigar cenários de possíveis discriminações aos usuários LGBTQIA+ que não foram punidas pela plataforma. Por fim, busca-se realizar um estudo comparativo da participação da população LGBTQIA+ em detrimento dos demais usuários (i.e., aqueles em que não há nenhuma autodeclaração).

Pacote de Replicação. Todos os *scripts* e dados utilizados nesta pesquisa encontram-se disponíveis em: https://doi.org/10.5281/zenodo.8178542

REFERÊNCIAS

[1] Mohammad Almarzouq, Abdullatif Alzaidan, and Jehad AlDallal. 2020. Mining GitHub for research and education: challenges and opportunities. *International Journal of Web Information Systems* ahead-of-print (06 2020). https://doi.org/10.1108/IJWIS-03-2020-0016

[2] Kelly Blincoe, Jyoti Sheoran, Sean Goggins, Eva Petakovic, and Daniela Damian. 2016. Understanding the popular users: Following, affiliation influence and leadership on GitHub. *Information and Software Technology* 70 (2016), 30–39. https://doi.org/10.1016/j.infsof.2015.10.002

[3] Riya Dutta, Diego Elias Costa, Emad Shihab, and Tanja Tajmel. 2023. Diversity Awareness in Software Engineering Participant Research. *SEIS - Software Engineering in Society* ahead-of-print (01 2023).

[4] Rita Garcia, Christoph Treude, and Wendy La. 2023. Towards Understanding the Open Source Interest in Gender-Related GitHub Projects. *arXiv preprint arXiv:2303.09727* (2023).

[5] GitHub Octoverse. 2022. Octoverse: Top programming languages. Website. https://octoverse.github.com/2022/top-programming-languages Acesso em: 24 de maio de 2023.

[6] Trysten Scott Richard, Eliane S Wiese, and Zvonimir Rakamarić. 2022. An LGBTQ-Inclusive Problem Set in Discrete Mathematics. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1*. 682–688.

[7] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. 2015. Gender and tenure diversity in GitHub teams. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3789–3798.

[8] Jennifer Wang and Sepehr Hejazi Moghadam. 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*. 615–620.