



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Os impactos da pandemia da Covid-19 nos projetos de software de código aberto

Marcio Vinicius Okimoto

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientadora
Prof.a Dr.a Edna Dias Canedo

Brasília
2021



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Os impactos da pandemia da Covid-19 nos projetos de software de código aberto

Marcio Vinicius Okimoto

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof.a Dr.a Edna Dias Canedo (Orientadora)
CIC/UnB

Prof. Dr. Uirá Kulesza
DIMAp/UFRN

Prof. Dr. Igor Fabio Steinmacher
Universidade Tecnológica Federal do Paraná (UTFPR)

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Programa de Pós-graduação em Informática

Brasília, 23 de novembro de 2021

Dedicatória

Essa pesquisa é dedicada a todas as famílias que perderam alguém durante a pandemia.

Agradecimentos

À minha família pela compreensão durante a minha ausência, enquanto me dedicava a esse trabalho. Aos professores Dra. Edna Dias Canedo e Dr. Rodrigo Bonifácio pelas orientações e ensinamentos ao longo da pesquisa.

Resumo

As medidas de isolamento social decorrentes da pandemia da COVID-19 alteraram as práticas de trabalho em diversos setores. Contudo, pouco se sabe sobre as implicações da pandemia nas atividades dos desenvolvedores de projetos de Software de Código Aberto (OSS). Nessa dissertação são analisados os efeitos da pandemia nas atividades de desenvolvimento de OSS e como o isolamento social mudou o estado emocional dos contribuidores. Para investigar esse problema, foi conduzido um método misto de pesquisa. Foi realizado um estudo quantitativo por meio da mineração do histórico de desenvolvimento de 155 repositórios; um estudo qualitativo que investigou as percepções dos desenvolvedores de OSS sobre o impacto da COVID-19, por meio de dois surveys; e uma análise de sentimentos em 2.661.886 comentários de pull requests para identificar mudanças no estado emocional dos contribuidores. Os resultados indicam que o advento da pandemia causou efeitos diversos nas atividades de desenvolvimento dos repositórios, especialmente no início das medidas de isolamento, como por exemplo, embora tenha sido observado um aumento no número de pull requests, também foi percebido um aumento na taxa de turnover dos core developers. Apesar disso, ao longo dos meses, houve uma estabilização desses indicadores. Outros efeitos foram mais abrangentes como a diminuição no ingresso de novos contribuidores. Alguns desses achados são suportados pelos resultados do estudo qualitativo, no qual a maioria dos participantes do questionário considera que a COVID-19 não mudou sua produtividade e entre os que os que perceberam alterações, houve uma dualidade na percepção. Esses achados podem ajudar a comunidade de OSS a compreender os efeitos da COVID-19 nas práticas de desenvolvimento e, dessa forma, subsidiar ações que possam mitigar os riscos e explorar as oportunidades, como a criação de políticas para aumentar o engajamento e reparar perdas criadas pela pandemia. Além disso, os resultados permitem que pesquisadores explorem outros pontos como o impacto a longo prazo dos efeitos da pandemia no processo de desenvolvimento de OSS.

Palavras-chave: Software Open Source, Mineração de Repositórios de Software, Github, Covid-19, Produtividade.

Abstract

The social isolation measures resulting from the COVID-19 outbreak changed work practices in various sectors. However, it is still unclear the implications of the pandemic on the daily activities of open source software (OSS) developers. In this study, we analyze the effects of COVID-19 on the development activity of OSS and how social isolation changed the productivity and emotional state of OSS contributors. To investigate this issue, we conducted a mixed-method study. We mined the development history of 155 open source projects; then we lead a qualitative study using two surveys as instruments in order to investigate the developers' perception about the impacts of the COVID-19 outbreak; and a sentiment polarity classification of 2,661,886 pull request comments in order to detect changes in the emotional state. Our results indicate that the pandemic situation led to some effect on the development activity of OSS repositories, especially in the early days of the outbreak, for instance, although we observe an increase in the number of pull requests, our results also indicate an increase in the turnover rate of core developers. Nonetheless, a few months later, the declaration of the pandemics situation, our results suggest stabilization in these metrics. Other effects were more lasting, like a significant decrease of newcomers along the pandemic period. Some of these findings are supported by the results of our qualitative study, whose results indicate that most of the respondents of our survey consider that the COVID-19 did not change their productivity substantially, but there is a duality between those who reported changes in productivity. Our findings can help the OSS community to understand the effects of the COVID-19 pandemic on development practices and in this way take actions to leverage the mitigation of risks and to explore opportunities. Some actions could be the creation of policies in order to increase engagement and reduce the impact of the pandemic. Furthermore, the findings can help researchers further investigate the effects of the pandemic on the OSS development process in the long term.

Keywords: Software Open Source, Mining Software Repository, Github, Covid-19, Productivity.

Sumário

| | | |
|----------|---|-----------|
| 1 | Introdução | 1 |
| 1.1 | Problema de Pesquisa | 2 |
| 1.2 | Justificativa | 2 |
| 1.3 | Objetivos | 3 |
| 1.3.1 | Objetivo Geral | 3 |
| 1.3.2 | Objetivos Específicos | 4 |
| 1.4 | Resultados Esperados | 5 |
| 1.5 | Contribuições do Trabalho | 5 |
| 1.6 | Metodologia de Pesquisa | 6 |
| 1.6.1 | Esforço de mineração | 6 |
| 1.6.2 | Survey | 7 |
| 1.7 | Estrutura do Trabalho | 8 |
| 2 | Indicadores em Repositórios de Software | 10 |
| 2.1 | Produtividade | 10 |
| 2.2 | Novos contribuidores e Contribuidores Casuais | 12 |
| 2.3 | Core Developers | 13 |
| 2.4 | Análise de sentimentos | 15 |
| 2.5 | Trabalhos Correlatos | 16 |
| 2.6 | Síntese do Capítulo | 20 |
| 3 | Configurações do Estudo | 21 |
| 3.1 | Esforço de Mineração | 21 |
| 3.1.1 | Etapa de Extração dos Dados | 24 |
| 3.1.2 | Etapa de Modelagem | 25 |
| 3.1.3 | Etapa de Síntese | 28 |
| 3.1.4 | Etapa de Análise | 28 |
| 3.2 | Survey | 31 |
| 3.3 | Síntese do Capítulo | 33 |

| | |
|--|-----------|
| 4 Resultados | 34 |
| 4.1 QP.1 - Qual o impacto da pandemia nas atividades de desenvolvimento nos repositórios de projetos de software de código aberto? | 34 |
| 4.2 QP.2 - Qual o impacto da pandemia no ingresso de novos contribuidores? . . | 44 |
| 4.3 QP.3 - Qual o impacto da pandemia no turnover de core developers? | 48 |
| 4.4 QP.4 - Qual a influência da pandemia no estado emocional e bem estar dos desenvolvedores de OSS? | 53 |
| 4.5 Síntese do Capítulo | 59 |
| 5 Análise dos Resultados | 60 |
| 5.1 Análises do esforço de mineração | 60 |
| 5.2 Análises dos surveys | 62 |
| 5.3 Discussões e Limitações | 64 |
| 5.4 Ameaças para Validação | 68 |
| 6 Conclusões | 70 |
| 6.1 Trabalhos Futuros | 71 |
| Referências | 72 |

Lista de Figuras

| | | |
|------|--|----|
| 1.1 | Metodologia da pesquisa, com o MSR e dois surveys. | 8 |
| 2.1 | Série Temporal - Esforço de Mineração dos Dados | 15 |
| 3.1 | Metodologia utilizada no processo de MSR. | 22 |
| 3.2 | 15 linguagens de programação com mais pull requests em março de 2020 | 24 |
| 4.1 | A mediana dos indicadores de atividade dos repositórios agrupados de forma mensal | 35 |
| 4.2 | Série temporal dos commits antes e depois da pandemia. | 37 |
| 4.3 | Série temporal do commit churn antes e depois da pandemia. | 38 |
| 4.4 | Série temporal do número de pull requests antes e depois da pandemia. | 40 |
| 4.5 | Série temporal da latência de pull requests antes e depois da pandemia. | 41 |
| 4.6 | Série temporal do número de review comments antes e depois da pandemia. | 42 |
| 4.7 | A mediana dos comentários dos repositórios agrupados de forma mensal | 43 |
| 4.8 | A percepção dos core developers sobre mudanças nos indicadores de atividade dos repositórios causados pela pandemia. | 44 |
| 4.9 | A média de novos contribuidores ao longo dos meses. | 45 |
| 4.10 | Ingresso de novos contribuidores ao longo dos anos. | 46 |
| 4.11 | Percentual de desenvolvedores que só fizeram um commit em relação a todos os novos contribuidores, ao longo dos meses. | 47 |
| 4.12 | Número de Newcomers vs. Newcomers que deixaram de contribuir com o projeto após a primeira contribuição. | 48 |
| 4.13 | Média da taxa de turnover de core developers, utilizando a contabilização de commits [1] % | 49 |
| 4.14 | Número de core developers que deixaram de contribuir com o projeto | 50 |
| 4.15 | Comparação do Turnover de Core Developers ao longo dos anos, utilizando a contabilização de commits [1] % | 51 |
| 4.16 | Comparação do turnover de core developers em relação a repositórios com diferentes níveis de atividade, utilizando a contabilização de commits [1] % | 52 |

| | |
|--|----|
| 4.17 Média da taxa de turnover de core developers, utilizando o algoritmo de TruckFactor % | 53 |
| 4.18 Mudanças na classificação da auto-produtividade dos participantes depois do início da pandemia, agrupados de acordo com o convívio familiar. | 55 |
| 4.19 Mudanças na classificação da auto-produtividade dos participantes, agrupados de acordo com o local de onde os participantes contribuía antes da pandemia. | 56 |
| 4.20 Sentimentos dos participantes durante as medidas de isolamento social. | 57 |
| 4.21 Como os participantes se sentiram durante o período de isolamento social. | 57 |
| 4.22 Polaridade dos sentimentos nos comentários de pull requests ao longo dos meses. | 59 |

Lista de Tabelas

| | | |
|-----|---|----|
| 1.1 | As métricas e as questões com a utilização do GQM | 4 |
| 3.1 | Estatística descritiva dos repositórios observados | 25 |
| 3.2 | Períodos observados e as variáveis temporais inseridas. | 29 |
| 4.1 | O efeito da pandemia no número de commits. A variável dependente é log(número de commits). $R_m^2 = 0.38$. $R_c^2 = 0.83$ | 37 |
| 4.2 | Repositórios que apresentaram diferença significativa em relação à distribui- ção do número de commits antes e durante a pandemia, divididos em quartis de acordo com o número de contribuidores. | 37 |
| 4.3 | Repositórios que apresentaram diferença significativa em relação à distribui- ção do número de commits antes e durante a pandemia, divididos em quartis de acordo com o número de linhas de código. | 38 |
| 4.4 | Repositórios que apresentaram diferença significativa em relação à distribui- ção do número de PRs antes e durante a pandemia, divididos em quartis de acordo com o número de contribuidores | 40 |
| 4.5 | O efeito da pandemia no número de pull requests. A variável dependente é o log(número de pull requests). $R_m^2 = 0.44$. $R_c^2 = 0.86$ | 41 |
| 4.6 | Diferença no ingresso de novos contribuidores depois do início da pandemia em relação à linguagem de programação usada nos repositórios | 47 |
| 5.1 | O resumo dos resultados do esforço de MSR e do Survey 2 em relação ao nível de atividade dos repositórios | 61 |
| 5.2 | Os resultados do Survey 1 e do Survey 2 em relação a percepção dos parti- cipantes sobre o nível de atividade dos repositórios. | 63 |
| 5.3 | Os resultados dos estudos relacionados ao desenvolvimento de software de código aberto. | 65 |

Capítulo 1

Introdução

No final do ano de 2019 houve o início da transmissão do novo coronavírus (COVID-19), inicialmente na cidade de Wuhan na China [2], fazendo com o que o mundo desse o primeiro passo a uma alarmante crise de saúde global que veio logo em seguida [3]. Com a rápida disseminação do vírus ao redor do mundo e um grande número de mortes [2], a Organização Mundial da Saúde (OMS) declarou, em 11 de março de 2020, a COVID-19 como uma pandemia [3]; muitos países encorajaram seus habitantes a realizar isolamento social e medidas legais foram adotadas em uma tentativa de mitigar a contaminação de pessoas [3]. Algumas dessas medidas foram: fechamento de escolas e universidades, fechamento de estabelecimentos que provêm serviços não essenciais, entre outros [3]. Com essas restrições, algumas empresas suspenderam suas atividades, enquanto outras decidiram adotar o Home Office [4], [5], [6].

O Home Office é uma modalidade onde se trabalha em uma localidade remota e, dessa forma, pode ser executada em qualquer lugar diferente de um escritório físico [7]. Contudo, ao invés de trabalhar em um escritório remoto ou em um lugar apropriado em casa, algumas pessoas passaram a trabalhar em lugares inadequados como quartos, mesas de cozinhas e sofás, enquanto parceiros, crianças, irmãos, pais, animais de estimação estão também no mesmo ambiente e podem ser uma fonte de distrações [6]. Além desses problemas, existe o stress associado com o fato de alguns profissionais terem contraído o vírus ou mesmo terem um parente em casa que necessite de cuidados [6]. Há também profissionais com filhos que tem que lidar com diversas tarefas relacionadas às aulas remotas, enquanto realizam suas atividades laborais [8].

O desenvolvimento de software de código aberto tem por natureza a característica de ser geograficamente distribuída [9] e, usualmente, os desenvolvedores de software não contribuem com projetos de código aberto como um trabalho formal [10]. Os resultados desse estudo indicam, inclusive, que a maior parte dos desenvolvedores (77,77%) já faziam a maior parte das suas contribuições para projetos de código aberto a partir de casa,

mesmo antes da pandemia. Contudo, as atividades dos contribuidores podem ter sido afetadas por mudanças em um emprego formal, ou por passarem a conviver com parceiros em Home Office ou com crianças em aulas remotas. Além disso, a pandemia por si só pode induzir a frustração, ansiedade e medo [6] e isso pode ter impactado as pessoas de uma maneira geral, inclusive os contribuidores de projetos de código aberto.

1.1 Problema de Pesquisa

Esse contexto encorajou pesquisadores a estudar como o isolamento social, durante a pandemia da COVID-19, está afetando a percepção de produtividade das pessoas [4] [6] , e o bem estar [11] , mas poucos estudos tiveram foco em verificar como o desenvolvimento de software de código aberto e seus desenvolvedores foram afetados pela pandemia [12] [13] [14]. Dessa forma, não se tem evidências suficientes se houve impacto no nível de contribuições dos repositórios de software de modo geral e no nível de atividades dos desenvolvedores de projetos de código aberto em função da pandemia.

Outro aspecto negligenciado no escopo dos projetos de código aberto é se a pandemia afetou a entrada de novos contribuintes e a manutenção dos chamados core developers (desenvolvedores que realizam a maior parte das tarefas relacionadas ao projeto [15]). Nesse caso, esses fatores têm o potencial de prejudicar o desempenho e a qualidade dos projetos [16] ou mesmo impactar a continuidade de projetos de código aberto [17]. Nesse contexto, as questões de pesquisa a serem respondidas nesse estudo são:

QP.1: Qual o impacto da pandemia nas atividades de desenvolvimento nos repositórios de projetos de software de código aberto?

QP.2: Qual o impacto da pandemia no ingresso de novos contribuidores?

QP.3: Qual o impacto da pandemia no turnover de core developers?

QP.4: Qual a influência da pandemia no estado emocional e bem estar dos desenvolvedores de OSS?

1.2 Justificativa

O processo de desenvolvimento de software de código aberto evoluiu muito nas últimas décadas, principalmente com a chegada de repositórios online como o GitHub ¹, que é uma plataforma online de colaboração, baseada no sistema de controle de versões git

¹<https://www.github.com>

[18]. Essa evolução levou os projetos de código aberto a serem vistos não apenas como alternativas viáveis, mas como soluções que alcançaram importância e sucesso na indústria de software [19], com diversos projetos sendo mantidos no GitHub. Entre esses projetos estão alguns bem populares como a biblioteca Javascript React ² ou o framework de Machine Learning Tensorflow ³. O GitHub, inclusive, é o maior repositório de software público existente atualmente [20], apenas em 2020 foram criados aproximadamente 60 milhões de novos repositórios [12], e a plataforma conta atualmente com mais de 56 milhões de desenvolvedores ativos [12].

Uma característica intrínseca aos projetos de código aberto é a criação e manutenção do código de forma colaborativa, permitindo times mais abertos e com maior diversidade, incluindo membros distribuídos geograficamente [20], ou seja, qualquer pessoa interessada, e possuindo os conhecimentos necessários, pode se envolver com o desenvolvimento dos projetos. Inclusive, o ingresso contínuo de novos colaboradores é crucial para o sucesso de projetos de código aberto [17].

Outro aspecto importante dos projetos de código aberto é o fato de que a parte mais significativa das contribuições é normalmente feita por um grupo restrito de desenvolvedores que possuem um longo período de engajamento com o projeto, esses desenvolvedores são conhecidos como core developers [21]. Segundo Mockus et al. [16] a saída desse tipo de contribuidor é o que mais impacta o desempenho e a qualidade dos projetos de software, em função da perda de conhecimento e experiência.

Dessa forma, dada a importância dos projetos de código aberto e dado o ecossistema complexo que envolve seu processo de desenvolvimento, torna-se importante investigar com maior nível de detalhes o impacto da pandemia nas atividades dos projetos, bem como no ingresso de novos contribuidores e na manutenção dos core developers.

1.3 Objetivos

1.3.1 Objetivo Geral

Essa dissertação possui dois objetivos principais. O primeiro objetivo é verificar se a pandemia causou impactos no processo de desenvolvimento de software de código aberto. Para alcançar esse objetivo, foi utilizado o método Goal-Question-Metric (GQM). Esse método emprega uma abordagem top-down e possibilita a determinação de objetivos e o refinamento dos objetivos em um conjunto de questões a serem respondidas. Essas questões são então desdobradas de forma que a coleta de dados específicos (métricas) possam responder as questões de forma quantitativa. As métricas podem ser objetivas,

²<https://github.com/facebook/react>

³<https://github.com/tensorflow/tensorflow>

quando dependem apenas do objeto a ser medido, ou subjetivas, quando elas dependem do objeto e do ponto de vista de quem observa. A Tabela 1.1 mostra as questões de pesquisa e as métricas associadas a cada uma delas. Inicialmente, são observadas métricas objetivas que detectam variações no nível de atividade dos repositórios, no ingresso de novos contribuidores e no turnover dos core developers. Adicionalmente, são observadas métricas subjetivas, sob o ponto de vista dos próprios contribuidores.

O segundo objetivo consiste em verificar o impacto da pandemia no bem-estar e estado emocional dos contribuidores. Nesse caso, inicialmente são utilizadas abordagens subjetivas, com a realização de estudos captando diretamente a percepção dos participantes dos projetos, com a própria percepção de produtividade, mudanças na motivação e a identificação dos principais sentimentos percebidos durante a pandemia. Posteriormente, é realizada a contabilização da polaridade dos sentimentos expressados nos comentários de pull requests, também com a intenção de identificar possíveis mudanças no estado emocional dos contribuidores.

| Questões | Métricas |
|---|--|
| Qual o impacto da pandemia nas atividades de desenvolvimento nos repositórios de projetos de software de código aberto? | <ul style="list-style-type: none"> • Número de commits • Número de pull requests • Latência de PRs • Commit churn • Número de comentários • Avaliação subjetiva da produtividade |
| Qual o impacto da pandemia no ingresso de novos contribuidores? | <ul style="list-style-type: none"> • Número de novos contribuidores • Taxa de contribuição casual |
| Qual o impacto da pandemia no turnover de core developers? | <ul style="list-style-type: none"> • Turnover mensal • Avaliação subjetiva do turnover |

Tabela 1.1: As métricas e as questões com a utilização do GQM

1.3.2 Objetivos Específicos

Para atingir o objetivo geral deste trabalho, os seguintes objetivos específicos foram definidos:

- Mensurar as diferenças entre as frequências de commits, pull requests, issues e comentários antes e durante a pandemia;
- Identificar os core developers e mensurar diferenças no turnover após o início da pandemia;
- Identificar os novos contribuidores e verificar variações no ingresso nos projetos, após o início da pandemia;
- Mensurar a percepção da auto-produtividade dos contribuidores durante o período de pandemia;
- Identificar os principais sentimentos dos contribuidores durante o período da pandemia;
- Analisar o bem estar dos contribuidores durante a pandemia.

1.4 Resultados Esperados

- Identificação de qual o efeito no nível de atividade dos contribuintes nos repositórios analisados após o início da pandemia;
- Verificação de alterações em tendências de ingresso e saída de core developers após o início da pandemia;
- Identificação dos níveis de novos contribuintes e de contribuições casuais antes e após o início da pandemia;
- Identificação de alterações na produtividade, motivação e no bem estar dos contribuintes durante o período de isolamento.

1.5 Contribuições do Trabalho

Este trabalho apresenta as seguintes contribuições:

1. Evidências sobre efeitos positivos e negativos da pandemia no nível de atividade dos repositórios, que podem ajudar a comunidade a entender essa situação inédita e a tomar medidas que possam potencializar os benefícios e mitigar os desafios associados à pandemia;

2. Identificação de alterações em aspectos relacionados a engajamento, especificamente o ingresso de novos participantes e turnover de core developers ao longo da pandemia. Fatores esses que tem o risco de prejudicar a execução dos projetos a médio e longo prazo. Esses dados podem inclusive lançar luz sobre a necessidade de se estimular o ingresso ou a retenção do conhecimento;
3. Identificação e caracterização da percepção dos desenvolvedores durante as medidas de isolamento decorrentes da pandemia, que podem auxiliar pesquisadores e profissionais a entender os principais sentimentos e o nível de bem estar dos desenvolvedores diante da nova realidade e a entender os impactos dessas mudanças no longo prazo.

1.6 Metodologia de Pesquisa

Este trabalho se baseia em um método misto com a utilização de estudos empíricos que combinam dados quantitativos e qualitativos. O estudo quantitativo é conduzido como um processo de Mineração de Repositórios de Software (MSR) [22] e tem por objetivo investigar as práticas de desenvolvimento e a estruturação do ambiente do GitHub [18]. Já o estudo qualitativo é realizado por meio de dois surveys com contribuidores de projetos de código aberto. O primeiro survey teve como público alvo todos os desenvolvedores que contribuíram com os projetos observados e o segundo survey teve como foco apenas os core developers dos projetos. Adicionalmente, técnicas de análise de sentimento são utilizadas para detecção de mudanças na proporção da polaridade expressos em artefatos de software. A Figura 1.1 ilustra a metodologia empregada: o primeiro quadro remete ao processo de MSR e a extração das métricas utilizadas para responder a QP.1 e para subsidiar as outras questões; com a identificação dos contribuidores o segundo quadro remete ao survey 1 e também à classificação da polaridade dos sentimentos nos comentários de pull requests utilizados para responder a QP.4; os contribuidores então são classificados em novos contribuidores (quadro 3) e core developers (quadro 4) e assim responder a QP.2 e QP.3. Por fim, o quadro 5 remete ao survey 2 utilizado para reforçar a resposta da RQ.1, com a triangulação dos dados do MSR.

1.6.1 Esforço de mineração

O desenvolvimento de software moderno envolve diversos atores que podem estar em localidades ou mesmo em fuso horários diferentes [23]. Dessa forma, tornou-se necessária uma plataforma para suportar a coordenação das atividades, bem como a comunicação dos diversos envolvidos no processo. Nesse contexto, surgiu o repositório de software GitHub,

a maior fonte de artefatos de software da Internet [18]. O processo de desenvolvimento no GitHub utiliza o conceito de "fork & pull", no qual desenvolvedores criam suas próprias cópias dos repositórios e, quando usam o paradigma do pull request, enviam uma solicitação de envio das mudanças para o mantenedor do projeto analisá-la e, se for o caso, aplicá-la no branch principal [18]. Além disso, a plataforma provê revisão de código, controle de issues e incorpora funcionalidades sociais [18].

Dessa forma, os repositórios de software passaram a conter informações muito valiosas sobre os projetos de software [24] e possibilitaram que os pesquisadores tivessem uma oportunidade única para observar o comportamento colaborativo dos desenvolvedores [25]. Inclusive, diversos estudos passaram a analisar os dados da plataforma, qualitativa e quantitativamente [18], usando a técnica de mineração de repositórios para se investigar práticas de desenvolvimento e a estruturação do ambiente do GitHub [18]. Nesta dissertação, é utilizado o guia proposto por Hemmati et al. [22], com as melhores práticas para um processo MSR, sendo o esforço de mineração baseado nos quatro temas do guia [22]: data acquisition and preparation, synthesis, analysis e sharing/replication. Além disso, o processo de MSR também é estruturado seguindo as quatro etapas sugeridas por Hemmati et al. [22]: extração, modelagem, síntese e análise.

Portanto, o objetivo do esforço de mineração nesta dissertação é extrair dados de commits, pull requests, issues e comentários; modelar core developers e novos contribuidores; sintetizar os dados em dois períodos, envolvendo meses antes e após o início da pandemia; e por fim analisar os efeitos da pandemia no nível de atividades dos repositórios.

1.6.2 Survey

Dois surveys foram executados a fim de se coletar informações qualitativas dos desenvolvedores de projetos de código aberto, capturando suas percepções sobre o impacto da pandemia da COVID-19 no nível de atividade dos repositórios, na percepção da produtividade e no bem-estar dos desenvolvedores. O primeiro survey foi baseado no estudo realizado por Ralph et al. [6], que investigou os efeitos da pandemia na produtividade e no bem-estar de desenvolvedores de software de empresas, não se atendo ao desenvolvedores de projeto de código aberto. Dessa forma, nesse estudo é realizada uma adaptação do questionário, removendo itens relacionadas ao vínculo empregatício, como questões que tem o foco no suporte oferecido por empresas aos desenvolvedores em trabalho remoto e adicionando itens relacionados à comunidade de código aberto. Esse estudo segue, portanto, a mesma metodologia no desenho do survey disponibilizado por Ralph et al. [6] em relação aos outros domínios investigados: o bem-estar emocional e a percepção da produtividade. O questionário está aderente ao *WHO's five-item wellbeing index* (WHO-5) para avaliação do bem estar, onde cada item é avaliado em uma escala de seis níveis indo de 0

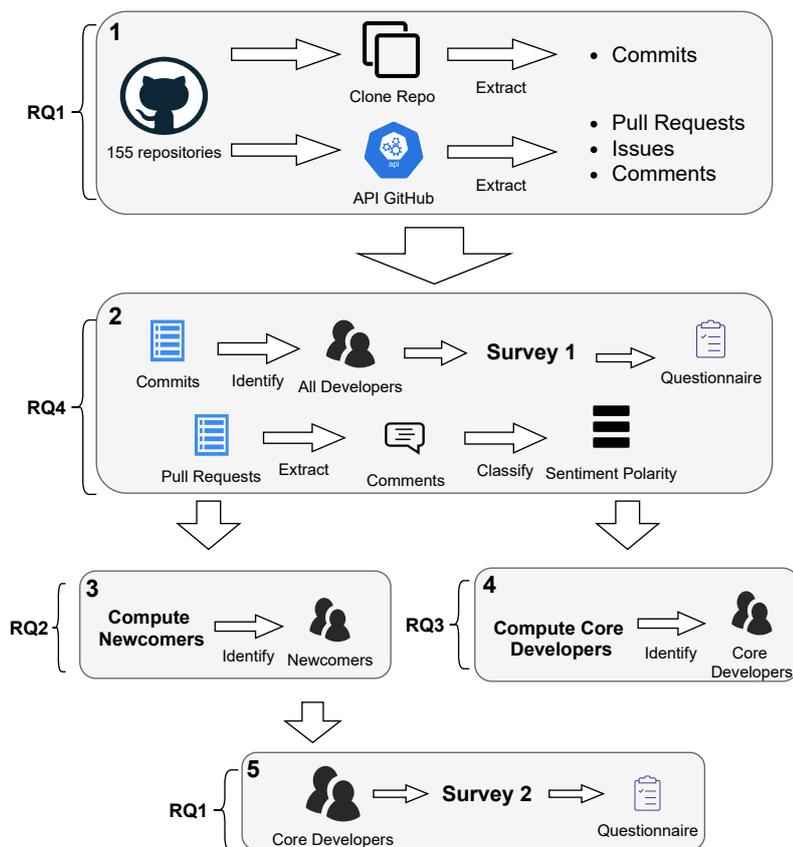


Figura 1.1: Metodologia da pesquisa, com o MSR e dois surveys.

a 5. E, ainda como no estudo original, os contribuidores avaliaram a auto produtividade duas vezes: antes e durante a pandemia. Mais detalhes sobre a configuração do survey serão apresentados no Capítulo 3. Já o segundo survey teve como público alvo apenas os core developers dos 155 projetos estudados e, tem como principal objetivo validar os achados do esforço de MSR.

1.7 Estrutura do Trabalho

Esta dissertação de mestrado está organizada em 6 capítulos, além deste, consistindo em:

- **Capítulo 2:** apresenta a fundamentação teórica necessária para o entendimento deste trabalho. Além disso, os trabalhos correlatos identificados na revisão de literatura são apresentados.
- **Capítulo 3:** descreve o método adotado para desenvolvimento deste trabalho e estabelece o processo de mineração nos repositórios de software realizado.

- **Capítulo 4:** demonstra os resultados obtidos pelas análises realizadas sob o ponto de vista do nível de atividade, taxa de entradas de novos contribuintes, do turnover dos core developers e expressa as análises qualitativas obtidas por meio do survey.
- **Capítulo 5:** apresenta a análise dos resultados e relaciona os resultados com os trabalhos correlatos.
- **Capítulo 6:** apresenta as principais conclusões deste trabalho e trata dos trabalhos futuros.

Capítulo 2

Indicadores em Repositórios de Software

Como a pandemia da Covid-19 é um fenômeno recente, porém de grande relevância, tendo impacto em vários níveis da sociedade e da economia, alguns estudos foram realizados para entender como a pandemia afetou o desenvolvimento de software. Alguns desses estudos investigaram o impacto na pandemia em grandes companhias como na Microsoft [26] ou na empresa Baidu [4]. Outros estudos tiveram uma abrangência mais ampla e realizaram pesquisas com desenvolvedores ao redor do mundo [11], ou em algumas regiões [27], [8] ou projetos específicos [5]. Alguns estudos [12], [13] [14] [28] lançaram luz sobre o impacto da pandemia na produtividade dos desenvolvedores e no nível e atividade dos projetos de software de código aberto, durante esse período da pandemia, mas esses estudos não abordaram alguns aspectos como o turnover dos core developers e o ingresso e engajamento de novos contribuidores.

No presente estudo são utilizadas técnicas já existentes, com algumas adaptações, para identificação dos novos contribuidores (newcomers) e dos contribuidores casuais (casual contributors). Também são identificados os core developers e seu turnover com heurísticas já existentes. As seções seguintes detalham os trabalhos que serviram de base para o cálculo desses indicadores. Por fim, são elencados os trabalhos correlatos.

2.1 Produtividade

A produtividade é essencial para o sucesso de um projeto de software [29] e o conceito de produtividade geralmente está ligado à relação entre entradas e saídas em um processo [30], como por exemplo o trabalho realizado em unidades de tempo [31]. Nesse sentido, algumas métricas foram utilizadas ao longo do tempo para se monitorar a produtividade em projetos de software de modo geral, tendo, inclusive, alguns estudos explorado mé-

tricas específicas de produtividade em repositórios de software [29]. Oliveira et al. [29] classificaram essas métricas em duas categorias: a primeira é baseada em características do código-fonte do projeto (code-based); já a segunda categoria tem métricas baseadas em commits (commit-based), que remete a características relacionada a commits realizados em um repositório.

Inclusive, as métricas baseadas em commits já foram usadas em estudos que investigaram como intervenções em projetos de código aberto impactaram a produtividade do time e dos indivíduos. Por exemplo, Vasilescu et al. [31] investigaram como a adoção do paradigma de Integração Contínua (CI) impactou a qualidade e a produtividade de projetos do GitHub. Os autores utilizaram as seguintes métricas para mensurar o impacto da produtividade: número de pull requests recebidos e o número de issues abertas durante os meses observados no estudo. Outro estudo relacionado aos impactos decorrentes da adoção de CI foi realizado por Zhao et al. [32]; mas nesse caso foram utilizados como indicadores do impacto da intervenção a média de commit churn, o número de issues abertas / fechadas, o número de pull requests, a média de pull requests abertos / fechados, a latência dos pull requests e o número de testes executado por build. Já o estudo de Wessel et al. [33] investigou o impacto da adoção de robôs para revisão automática de pull requests; para análise do efeito foram utilizados indicadores como número de pull requests aceitos e não aceitos, número de comentários, latência de pull requests e número de commits.

Além disso, métricas baseadas em commits também já foram usadas em estudos que investigaram questões ligadas à diversidade dos times de projetos de código aberto, como no estudo de Vasilescu et al. [20] em que foi empregado o número de commits para mensuração da produtividade. Outro estudo na mesma linha foi realizado por Ortu et al. [34], mas nesse caso foi utilizado como um indicador de produtividade apenas o tempo para se resolver issues.

Em relação a estudos específicos da pandemia, os relatórios lançados pelo próprio GitHub [13], [12] investigaram o fenômeno no âmbito de projetos de código aberto. Quando analisada a produtividade individual dos desenvolvedores, o estudo observou as seguintes métricas: número de pull requests, pushes, revisões de pull requests e comentários de issues por contribuinte. Já o estudo de Silveira et al. [14] utilizou como métricas do nível de atividade dos repositórios: o número de branches, pull requests abertos e fechados, número de commits e número de comentários. Por fim, Wang et al. [28] utilizaram métricas baseadas nas atividades dos usuários, como abertura de issues, pull requests submetidos, etc.

Outra forma de se aferir a produtividade dos indivíduos é perguntando diretamente a eles suas percepções. Embora não se tenha um consenso sobre qual a melhor forma seja a

melhor (a utilização de métricas objetivas ou a autoavaliação), estudos foram realizados para entender quais práticas mais afetam a autoavaliação de desenvolvedores [35], [36] e concordam no fato de que a percepção da atividade é pessoal e varia entre os indivíduos.

No presente estudo são empregados indicadores baseados em commits para mensuração da produtividade dos times, uma vez que o commit é um fator abrangente da representatividade de produtividade em projetos de código aberto [20]. Dessa forma são observados o número de commits abertos por mês e a mediana do número de commit churn de cada repositório por mês, antes e durante a pandemia. Adicionalmente, são definidos indicadores relacionados ao processo de pull requests: número de pull requests abertos por mês e a mediana da latência dos pull requests de cada repositório por mês. A adição desses indicadores tem o intuito de se ter uma visão mais ampla da produtividade dos times, pois o processo de pull request usualmente envolve uma série de passos e ao menos dois atores [37]. No caso dos surveys, é utilizada a autoavaliação da produtividade, questionando os participantes sobre qual sua percepção em uma escala de 0 a 10, antes e durante o período de isolamento. Essa mesma escala foi utilizada no estudo realizado por [6] que por sua vez foi baseada no *WHO's Health at Work Performance Questionnaire (HPQ)*¹.

2.2 Novos contribuidores e Contribuidores Casuais

A interrupção do fluxo de entrada de novos contribuintes tem o potencial de ameaçar a continuidade dos projetos [17]. Além disso, a entrada de novos contribuidores e seu engajamento ativo é importante para o sucesso de projetos de código aberto [38], dessa forma, diversos estudos foram realizados para entender os motivos pelos quais esses novos contribuintes desistem de participar dos projetos [39],[40], [41]. Dada a importância desse tipo de desenvolvedor, é importante verificar se houve variações nesse fluxo também ao longo do período de pandemia. Nessa dissertação a taxa de ingresso de novos contribuidores é calculada verificando o primeiro commit realizado nos projetos do nosso dataset, sendo essa verificação também utilizada no estudo realizado por Bao et al. [42].

Contribuidores casuais (também conhecidos como drive-by-commits) são desenvolvedores que fazem contribuições, mas não tem um envolvimento de longo prazo com o projeto [43]. Embora normalmente contribuam com uma pequena parte do total de contribuições [43],[44], muitas vezes essas contribuições não são triviais, pois comumente envolvem bug fix, adição de novas features e refatoração de código [43]. Nesse sentido, torna-se importante verificar também se a pandemia afetou as contribuições casuais. Nessa dissertação é utilizada a abordagem definida por Pinto et al. [43] para a identificação de contribui-

¹<https://www.hcp.med.harvard.edu/hpq/info.php>

dores casuais: desenvolvedores que fizeram no máximo uma contribuição durante toda a existência do projeto.

Por fim, com o intuito de se investigar mudanças no comportamento desse desenvolvedores, são utilizadas as métricas taxa mensal de ingresso de novos contribuidores nos projetos e taxa mensal de contribuições casuais em relação ao total de commits.

2.3 Core Developers

Core developers são aqueles desenvolvedores que contribuem significativamente com o desenvolvimento de um sistema [15], que exercem um papel essencial no desenvolvimento da arquitetura de um sistema [45] e são responsáveis pelas tarefas mais importantes em um projeto de software como no design, implementação e manutenção [1]. Nesse sentido, a continuidade dos projetos podem ser comprometidas, caso os core developers decidam não mais contribuir com o desenvolvimento do sistema [15].

No contexto de projetos de software de código aberto, algumas abordagens foram sugeridas para classificar contribuintes em core developers de forma quantitativa: algumas delas contabilizam a quantidade de linhas de código que o desenvolvedor enviou para o projeto [46], enquanto outras utilizam a contabilização de commits [1] ou contabilização da taxa de arquivos que um dado desenvolvedor é considerado dono [47]. Nesse último caso é utilizada uma heurística para atribuição da propriedade do arquivo a um desenvolvedor específico, que inclusive pode mudar ao longo do tempo. De forma contrastante, Joblin et al. [45] propuseram uma abordagem network-based, que não leva em conta apenas count-based metrics, como o número de linhas ou o número de commits, mas aspectos qualitativos envolvidos no processo de desenvolvimento do software.

Ferreira et al. [1] abordaram o fenômeno do turnover de core developers em repositórios do GitHub e concluíram que repositórios com alta taxa de turnover levam um maior tempo para encerrar issues e consertar bugs, bem como para implementar melhorias. Dessa forma, é importante verificar se a pandemia impactou a permanência dos core developers nos projetos de software de código aberto (OSS) e, conseqüentemente, se houve impacto na continuidade dos projetos OSS.

Nesse dissertação, são utilizadas duas abordagens para identificação dos core developers. A primeira é a mesma heurística utilizada por Ferreira et al. [1] para definição dos core developers: desenvolvedores que foram autores de 80% dos commits dos repositórios, excluindo desenvolvedores que contribuíram com um valor menor que 5% do total dos commits do período. A segunda abordagem utiliza o algoritmo de Truck Factor [47], que foi executado contra os repositórios com o auxílio da ferramenta automatizada disponibi-

lizada por Avelino et al. [47]. A partir do resultado das duas abordagens, foi calculado o turnover para se ter uma análise mais ampla em relação a esse tipo de desenvolvedor.

Em relação ao cálculo do turnover dos core developers, foi utilizada a mesma abordagem de Ferreira et al. [1], mas com algumas adaptações. O método utilizado no estudo de Ferreira et al. [1] realizou o cálculo do turnover comparando os core developers de um ano com o do ano imediatamente anterior. De forma semelhante, nessa dissertação são computados os core developers levando em conta as contribuições realizadas no período de um ano; contudo, para efeitos de comparação, o turnover é calculado em bases mensais. Por exemplo, para calcular a taxa de turnover de maio de 2020, primeiro são computados os core developers dos 12 meses anteriores (nesse caso, o período de maio de 2019 a abril de 2020) e eles então inseridos, por exemplo, em um conjunto chamado SET A; em seguida são computados os core developers dos 12 meses que incluem maio de 2020 (ou seja, o período de junho de 2019 a maio de 2020), com os desenvolvedores sendo incluídos em um conjunto chamado SET B; e por fim, as taxas de entrada e de saída de core developers, necessária para o cálculo do turnover, são realizadas de acordo com as fórmulas do estudo de Ferreira et al. [1].

Inicialmente, são calculadas as taxas de desenvolvedores que passaram a ser Core Developer (CDN) no período, utilizando a fórmula [1] abaixo:

$$CDN_{Rate}(n) = \frac{|CDN(n)|}{|SETB|} \times 100$$

Sendo:

$$CDN(n) = SET B - SET A$$

Em seguida é calculada a taxa de desenvolvedores que deixaram de ser core developers (CDL) em cada um dos meses observados. Nesse caso, são os desenvolvedores que são core developers e não contribuíram o suficiente para se manter como core developers no período posterior.

$$CDL_{Rate}(n) = \frac{|CDL(n)|}{|SETA|} \times 100$$

Sendo:

$$CDL(n) = SET A - SET B$$

Por fim, é calculada a taxa de Turnover (CDT Rate) também de acordo com a fórmula de Ferreira et al. [1].

$$CDT_{Rate}(n) = \frac{|CDL(n)|}{avg(|SETA| + |SETB|)} \times 100$$

Importante notar que a entrada de novos core developers ajuda a amenizar a taxa de turnover, dado que os novos core developers estão no SET B, conforme apresentado na Figura 2.1.

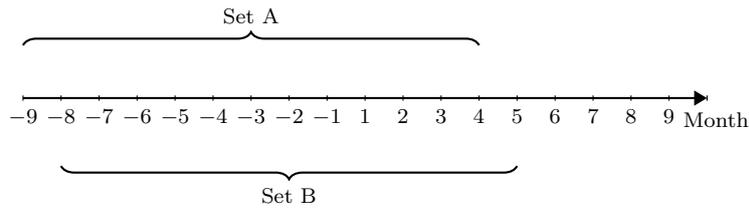


Figura 2.1: Série Temporal - Esforço de Mineração dos Dados

Dessa forma, nessa dissertação são utilizadas as taxas mensais de: entrada, saída e turnover de core developers nos repositórios.

2.4 Análise de sentimentos

Alguns trabalhos estudaram a análise de sentimento no escopo da engenharia de software em questões relacionadas ao ambiente do GitHub [48] [49] [50]. Guzman et al.[48] analisaram comentários de mais de 29 projetos escritos em 14 diferentes linguagens de programação. Adicionalmente, foi feita uma análise para identificar a quais fatores as emoções detectadas estão relacionadas. Para classificação da polaridade dos sentimentos (positivo, neutro ou negativo) foi utilizada uma abordagem lexical com o emprego da ferramenta SentiStrength [51]. Os resultados demonstram uma presença de mais comentários negativos em projetos escritos em Java e uma maior frequência de comentários positivos em times mais distribuídos. Além disso, foi encontrada uma maior proporção de sentimentos negativos quando escritos nas segundas-feiras.

Outro estudo que explora sentimentos em comentários do GitHub foi conduzido por Pletea et al. [49]. Nesse caso, são estudados aspectos relacionados a discussão em torno de questões de segurança em 90 projetos do GitHub. Os autores classificaram a polaridade dos sentimentos dos commits utilizando a ferramenta de propósito geral chamada NLTK [52], que foi treinada com um dataset de filmes. Como resultado, concluíram que discussões relacionadas a questões de segurança implicam uma maior quantidade de sentimentos negativos em relação a outros tipos de discussões.

O foco do estudo de Sinha et al. [50] também foi entender as emoções dos contribuidores em 2.130.474 commits de repositórios do GitHub. Foi utilizada como ferramenta de classificação de sentimentos o SentiStrength [51], ferramenta desenhada para classificar mensagens do Twitter. Como resultado, os autores encontraram uma maior frequência de sentimentos negativos em projetos com mais commits. Além disso, foi encontrada uma maior frequência de sentimentos negativos em commits realizados nas terças-feiras e maior frequência de sentimentos positivos nas sextas-feiras. Os autores também encontraram uma correlação entre os sentimentos e o número de arquivos alterados nos commits.

Algumas ferramentas de análise de sentimentos foram criadas especificamente para o contexto da engenharia de software, tais como: SentiStrength-SE [53] e DEVA [54] que utilizam uma abordagem lexicon-based. As ferramentas Senti4SD [55] e SentiCR [56] utilizam um modelo supervisionado, com o treinamento prévio para a classificação da polaridade dos textos. Em um estudo anterior Novielli et al. [57] compararam algumas dessas ferramentas e demonstraram que implementações específicas para o domínio de engenharia de software têm resultados superiores às ferramentas de propósito geral; sendo que o melhor resultado foi apresentado por ferramentas que implementam abordagens supervisionadas. Assim, nessa pesquisa utilizamos a ferramenta de classificação de sentimentos Senti4SD com o treinamento utilizando um dataset público [58] com textos do GitHub e anotações quanto à polaridade dos sentimentos. A escolha foi baseada no fato de que trata-se de uma ferramenta que utiliza a abordagem de treinamento supervisionado. Além disso, evidências empíricas [39] demonstraram a superioridade dessa ferramenta ao classificar textos do GitHub em relação a outras ferramentas específicas para o âmbito da engenharia de software.

2.5 Trabalhos Correlatos

Bao et al. [4] realizaram uma investigação sobre as diferenças entre o trabalho de casa (WFH) e o trabalho tradicional face a pandemia da Covid-19. Com esse objetivo, foi feita uma análise quantitativa baseado em um dataset que continha as atividades diárias de um grupo de desenvolvedores da empresa Baidu Inc ². Os dados coletados para o estudo tinham registros de atividades de 139 desenvolvedores em 138 dias de trabalho. As métricas utilizadas na pesquisa foram: número de builds, commits e revisões de códigos. Os resultados mostraram que o WFH tem impactos positivos e negativos na produtividade e que esses impactos são diferentes dependendo das características dos projetos em que os desenvolvedores trabalham, como por exemplo a linguagem de programação usada.

A fim de entender o impacto da pandemia nos desenvolvedores, o estudo de Ralph et al. [6] gerou e avaliou um modelo teórico para explicar e prever mudanças no bem estar e produtividade em relação ao trabalho em casa durante a crise da COVID-19. Para conduzir a análise comportamental durante a pandemia, os autores usaram um survey, traduzido em 12 idiomas diferentes, e submetido para diversos profissionais na área de desenvolvimento de software. Como contribuição, o trabalho ofereceu recomendações aos profissionais e empresas para darem o suporte para pessoas em trabalho remoto durante a pandemia e para futuras crises.

²<https://www.baidu.com/>

Seguindo a metodologia de análise exploratória, Ford et al. [59] distribuíram dois surveys entre funcionários da Microsoft e obtiveram 3.634 respostas. Os autores queriam entender a presença e prevalência de benefícios, desafios e oportunidades para melhorar essa situação atípica de trabalho remoto em função da pandemia. Os autores usaram análises qualitativas e quantitativas e descobriram que um mesmo fator pode ter impactos positivos e negativos nas experiências dos desenvolvedores, como por exemplo, estar próximo da família pode aumentar o bem estar dos participantes, mas também pode ser uma fonte de interrupções. Em relação à produtividade, nos dois surveys os participantes foram indagados sobre diferenças na sua própria produtividade e os resultados indicaram que a maior parte dos participantes mantiveram o mesmo nível ou aumentaram a produtividade (62% e 68%, respectivamente). Como forma de triangular o nível de produtividade, adicionalmente eles verificaram métricas relacionadas ao processo de desenvolvimento: eles compararam a quantidade de pull requests por desenvolvedor no período de pandemia com períodos anteriores, e concluíram que não houve mudanças significativas na produtividade.

Outro estudo exploratório foi conduzido por Lal et al. [60], usando a técnica de escrita de um diário com 29 participantes. Os participantes foram estimulados a completar um diário com registros em um período de 10 dias úteis com questões relacionadas a aspectos emocionais e sociais. Os resultados indicaram que a interação social está em declínio para a maior parte dos trabalhadores em trabalho remoto, mas apenas um terço relataram sentirem falta de interações sociais com colegas no local de trabalho.

Bezerra et al. [8] pesquisaram como os fatores humanos e organizacionais impactaram na produtividade de desenvolvedores brasileiros durante a pandemia, por meio de um survey com 58 participantes. De forma interessante, a maioria das pessoas que responderam ao questionário (74.1%) reportaram que a produtividade não diminuiu durante o período de pandemia e 84.5% deles se sentiram motivados e até perceberam um aumento na comunicação com seus colegas de trabalho.

De uma forma mais ampla, Russo et al. [11] executaram um estudo longitudinal em duas etapas, envolvendo 200 profissionais de software a fim de caracterizar um dia típico de trabalho de um desenvolvedor durante a pandemia e os impactos das atividades individuais no lado sentimental e no bem estar dos trabalhadores. Os resultados indicaram que não houveram diferenças significativas relacionadas à distribuição do trabalho, embora eles tenham percebido que desenvolvedores gastam menos tempo em reuniões e pausas durante a pandemia. Interessante citar que não foram encontradas relações diretas entre atividades específicas e a produtividade/bem-estar dos desenvolvedores.

Explorando posts em redes sociais, Xiong et al. [61] analisaram tweets de 58.345 pessoas dos Estados Unidos no Twitter users durante a pandemia e que mencionaram

hashtags ou keywords relacionados ao WFH. Os autores reportaram que o sentimento do WFH variou dependendo das características dos usuários. De forma mais específica, segundo os autores, mulheres, pessoas na faixa dos 30 anos e em áreas urbanas tiveram um sentimento positivo relacionados ao WFH. O estudo analisou profissionais de uma forma geral e não teve foco apenas em desenvolvedores de software.

A experiência de gerenciar um projeto durante a pandemia do COVID-19, foi reportada por Mendonça et al. [5]. Eles demonstraram as mudanças em práticas requeridas e os riscos envolvidos durante o período de adaptação para as medidas de isolamento da pandemia da COVID-19. De acordo com os autores, essas mudanças não afetaram negativamente o nível de produtividade. Além disso, eles apresentaram as percepções dos stakeholders e do time de desenvolvimento sobre o impacto da pandemia no projeto. Nesse sentido, o maior desafio reportado está relacionado à disponibilidade e comunicação com o time de stakeholder durante as fases do projeto.

O principal objetivo da pesquisa conduzida por Garro-Abarca et al. [62] foi verificar o relacionamento de fatores relacionados a times virtuais encontrados previamente na literatura. Foram analisadas o desempenho de 317 engenheiros de software em times virtuais. Os resultados sugerem que a confiança é a base para a maior parte das variáveis analisadas e que esse fator deveria ser levado em consideração em todas as medidas relacionadas a liderança, comunicação, coesão ou configuração de características das tarefas em times distribuídos remotamente.

Com dois surveys realizados com engenheiros de software em uma grande companhia localizada nos Estados unidos, Miller et al. [26] investigaram como o WFH em função da pandemia da COVID-19 afetou a cultura e a produtividade das equipes. O primeiro survey foi conduzido para se entender o impacto do WFH e os resultados iniciais mostraram que muitos desenvolvedores encontraram desafios para atingir os milestones e que a comunicação e a interação social foram afetadas. O segundo survey, planejado a partir dos impactos do primeiro, demonstraram que 74% das pessoas que responderam sentiram falta da interação social com os colegas e que 51% reportaram um aumento na dificuldade de comunicação com os colegas. Um modelo de regressão demonstrou que a comunicação, a habilidade de brainstorming com os colegas e as interações sociais são os fatores de cultura mais importantes relacionados à produtividade.

A pandemia estimulou uma quantidade enorme de pesquisas por parte de cientistas em várias disciplinas, pesquisando pelo fenômeno em si, sua característica epidemiológica e meios para se enfrentar suas consequências [63]. Georgiou et al. [63] investigaram a extensão, evolução e características de posts relacionados com a COVID-19 em posts da plataforma Stack Overflow ³. Outro estudo nessa linha foi realizado por Rahman et

³<http://stackoverflow.com/>

al. [64], no qual os autores caracterizaram projetos criados no GitHub para analisar e mitigar as consequências da COVID-19. Os autores também caracterizaram e quantificaram os bugs relacionados a esses projetos. Já Oliveira et al. [65] analisaram projetos do GitHub e perguntas do Stack Overflow relacionados a COVID-19 a fim de caracterizar o desenvolvimento de softwares relacionados com a pandemia.

Outros estudos também foram realizados para entender o impacto da pandemia da COVID-19 em equipes que utilizam metodologias ágeis [66, 27, 67]. Nesse sentido, o estudo de Marek et al. [67], que executou um survey global com 120 participantes, concluiu que não houve impacto na produtividade das equipes. Na mesma linha, o estudo de Neumann et al. [27], que analisou dados de equipes ágeis localizadas na Alemanha, também reportou que a maioria dos times não perceberam nenhuma perda de desempenho. Já o trabalho de Butt et al. [66] reportou que a indústria de software usando metodologias ágeis durante a pandemia tiveram um impacto negativo, com redução de 82% na influência dos processos ágeis.

No âmbito dos projetos de software livre, relatórios do próprio GitHub trataram do impacto da pandemia [13] [12]. O GitHub Octoverse 2020 Productivity Report [12] investigou a cadência de trabalho e as atividades de repositórios privados e públicos, durante o ano de 2020. Inicialmente, foram analisados dois indicadores: push windows (tempo entre o primeiro e o último push do dia) e o push volume (representado pelo número de pushes em um dia). A análise dos dados mostrou um aumento no nível de atividade durante os feriados e finais de semana em relação ao ano anterior à pandemia. Também foi observada uma tendência dos desenvolvedores contribuírem mais com projetos de código aberto durante os finais de semana, o que sugere que desenvolvedores full-time utilizam os finais de semana para contribuir para projetos de lazer ou para fins educacionais. De uma maneira geral, houve um aumento do nível de atividade no começo das medidas de isolamento social, com uma estabilização ao longo dos meses de 2020. Também foram investigadas as atividades dos desenvolvedores, por meio dos indicadores pull requests, pushes, revisões de pull requests, e comentários em issues por pessoa. Os resultados indicam que o nível de atividade dos repositórios são equiparados com o nível do ano anterior à pandemia.

Um estudo preliminar teve foco nos impactos da pandemia no nível de contribuições dos desenvolvedores. Conduzido por Wang et al. [28] foram analisadas métricas raramente utilizadas na literatura, como a abertura de issues e a submissão de pull requests. Os autores afirmam terem coletado dados de todos os desenvolvedores do GitHub e analisado esses dados sob as perspectivas de ecossistemas, país, da organização e do desenvolvedor. O resultado do estudo indicam um aumento na soma de contribuições e no número de desenvolvedores ativos no início da pandemia; que o aumento dessas contribuições podem estar correlacionados com o surto de COVID-19 nos países dos desenvolvedores; que

não foi encontrada diferenças entre desenvolvedores contratados por grandes companhias e que grupos de contribuidores baseado em níveis de atividade apresentaram diferentes tendências durante a pandemia.

O estudo de Silveira et al. [14] também teve foco nos projetos de software livre, nesse caso foram observados os históricos de contribuições de 100 repositórios escritos na linguagem Java e um survey com 279 profissionais. Foram realizadas 12 observações que traçam um panorama do impacto da COVID-19. Como conclusão, os autores realtam que os impactos da COVID-19 não é binário (com redução ou aumento da produtividade), sendo que diferentes contribuidores tem visões diferentes de um mesmo aspecto.

Outro estudo investigou os efeitos na pandemia no GitHub e no Stackoverflow [68]. Em relação ao resultados do GitHub, os autores não conseguiram concluir com certeza se a pandemia teve um efeito no aumento de atividades; contudo foram identificados alguns desenvolvedores com um aumento expressivo na frequência durante os meses de março e abril de 2020 (que abrangem o início das medidas de isolamento social em todo o mundo).

Diferentemente da maioria dos estudos correlatos apresentados, o presente estudo investigou especificamente o desenvolvimento de software de código aberto. Além disso, esse estudo apresenta diferenças em relação aos estudos que focaram nos efeitos da pandemia em OSS, uma vez que fatores adicionais foram abordados, tais como: o impacto no ingresso de novos participantes e o turnover dos core developers face o advento da pandemia. Além disso, outro diferencial é que o método aplicado abrange análises não usadas anteriormente, como por exemplo o uso do RDD.

2.6 Síntese do Capítulo

Este capítulo apresentou conceitos necessários para a contextualização desta dissertação. Nesse sentido, foi apresentado o conceito de produtividade e os indicadores mais utilizados para sua medição na literatura. Também foram descritos os conceitos de Core Developer e casual developers, bem como abordadas as heurísticas utilizadas para identificá-los. Além disso, foi descrita a metodologia utilizada nos dois estudos realizados: a mineração de repositórios e os surveys. Por fim, foram apresentados trabalhos que tiveram como objetivo analisar os efeitos da pandemia da COVID-19 em diversas áreas.

Capítulo 3

Configurações do Estudo

Este Capítulo apresenta como foi estabelecido o esforço de mineração dos repositórios de software e os surveys com os contribuidores. O objetivo desse estudo é prover evidências que possam demonstrar os possíveis efeitos da pandemia na manutenção dos projetos e também na identificação das causas relacionadas a esses impactos. O pacote de reprodução¹ para o esforço de Mining Software Repositories (MSR) e para os surveys está disponível na plataforma de arquivamento Zenodo.

3.1 Esforço de Mineração

Com o objetivo de responder as questões de pesquisa QP.1, QP.2 e QP.3, um esforço de mineração foi conduzido de forma a permitir a investigação do impacto da pandemia no desenvolvimento de projetos de software de código aberto sob três perspectivas: (i) o nível geral de atividade dos repositórios no período; (ii) a entrada e a relevância das contribuições de novos contribuidores no nível de atividade dos repositórios; e (iii) o turnover dos contribuidores mais experientes (core developers) . A Figura 3.1 ilustra a metodologia utilizada no processo MSR. O processo foi baseado nas recomendações de Hemmati et al. [22], sendo dividido em 4 temas: data acquisition, synthesis, analysis e sharing/replication.

O tema data acquisition compreende as etapas de **Extração** e **Modelagem** dos dados [22]. No caso do Github os dados podem ser extraídos de duas formas [25]. Na primeira abordagem é utilizada uma forma mais acessível e padronizada, como os bancos de dados públicos GHTorrent [69] - nesse caso há restrições, como a disponibilização de apenas um subconjunto dos dados [25], o que, a depender do tipo de investigação a ser realizada, pode inviabilizar as análises. A outra possibilidade é a extração ser feita de forma manual com a realização de clones ou com o uso da API do GitHub. No caso da utilização de clones, os

¹<https://zenodo.org/record/4771078#.YKQqRqhKiiM>

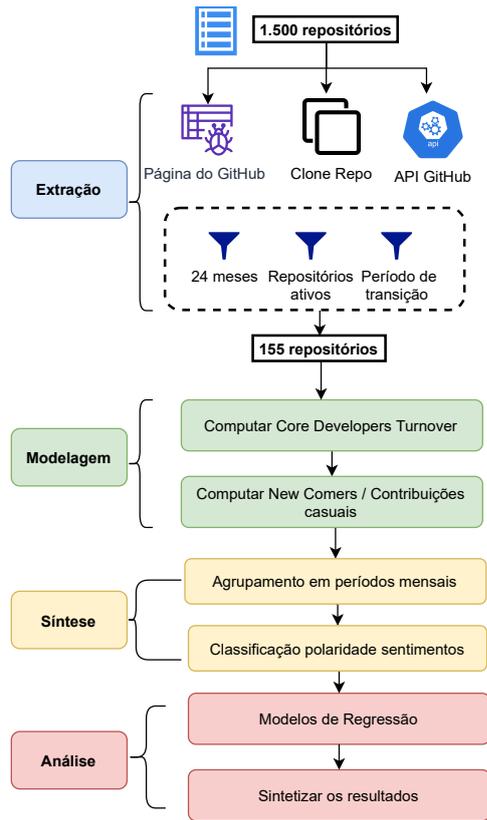


Figura 3.1: Metodologia utilizada no processo de MSR.

utilitários git devem ser usados para extração dos dados, especialmente o git log, que provê versatilidade e possibilita a visualização de metadados sobre os commits [25]. No caso da utilização das APIs, há questões relacionadas aos limites impostos pelo mantenedor do serviço, nesse caso deve-se frequentemente verificar se o limite não foi atingido [25], sob risco do acesso ser negado. Nesse estudo é utilizada a extração manual dos metadados, com clone dos repositórios para coleta dos dados dos commits e a API do GitHub para extração de dados relacionados a pull requests. Adicionalmente, são utilizadas técnicas de web scraping para obtenção de informações não disponíveis em outras fontes - essa técnica utiliza a extração dos dados diretamente das páginas do GitHub, sendo permitida pelos termos de uso do GitHub [25]. O web scraping é aconselhado nos casos em que a informação não pode ser obtida por meio de outras fontes e também como forma de contornar as limitações impostas pela API [25]. Em relação à modelagem, também são utilizadas as recomendações de [22] para agrupamento dos dados das diversas fontes e para a modelagens dos desenvolvedores.

O tema synthesis, compreende a fase onde tipicamente os dados das fases anteriores são agrupados, classificados, sofrem algum tipo de predição ou passam por algum outro

algoritmo de machine learning [22]. Nesse estudo, os dados são agrupados em períodos mensais e observadas as recomendações de Hemmati et al. [22] quanto à necessidade de se atentar à natureza da distribuição dos dados em relação à normalidade para escolha do método estatístico a ser usado. Outra recomendação seguida a partir das boas práticas sugeridas por Hemmati et al. [22] é a utilização da escala logarítmica nos modelos de regressão onde as variáveis dependentes ou independentes possuem distribuição não-normal.

Já no tema analysis, Hemmati et al. [22] lista as boas práticas referentes à etapa onde os resultados são analisados e interpretados. É visto como a fase mais importante em um processo de MSR [22], uma vez que as conclusões são resultados diretos da saída desta fase. Nesse sentido, torna-se necessária a utilização de métodos estatísticos rigorosos, verificação manual e a utilização de técnicas adequadas de visualização [22] para se ter resultados corretos. Durante as análises desse estudo são aplicadas as boas práticas sugeridas por Hemmati et al. [22] em relação aos testes estatísticos e às medidas de tendência central a serem utilizadas.

Por fim, no tema sharing/replication são feitas considerações sobre como os dados devem ser armazenados e compartilhados para que seja possível ser feita uma validação externa ou replicação do estudo [22]. A fim de tornar os dados públicos e disponíveis, nessa dissertação as etapas foram detalhadas ao máximo e compartilhadas em plataformas de arquivamento, que garantem que os dados possam ser consultados no futuro, evitando links quebrados ou outras indisponibilidades dos dados. Também são disponibilizados em repositório público os scripts de todas as fases do processo de MSR, sendo totalmente possível a reprodução completa dessa dissertação, incluindo a extração e modelagem dos dados dos repositórios via clone e API do GitHub, bem como a execução da síntese e análise dos dados. Ainda seguindo a recomendação de Hemmati et al. [22], os dados brutos e processados foram disponibilizados.

Dessa forma, foram necessários diversos passos que serão descritos nas seções seguintes para obtenção, modelagem, síntese e análise dos dados. Além disso, com o intuito de tornar os dados públicos e disponíveis para reprodução, as etapas desse processo foram detalhadas e compartilhadas em plataformas de arquivamento, que garantem que os dados possam ser consultados no futuro, evitando links quebrados ou outras indisponibilidades dos dados [22]. Também foram disponibilizados em um repositório público os scripts de todas as etapas da mineração, sendo possível a reprodução completa dessa dissertação. Seguindo a recomendação de [22], os dados brutos e processados foram disponibilizados.

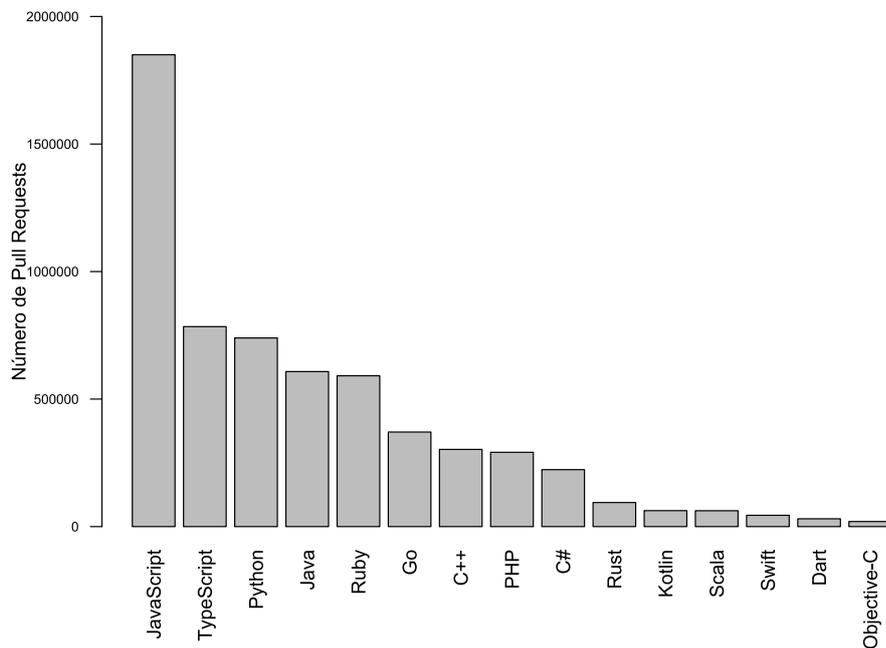


Figura 3.2: 15 linguagens de programação com mais pull requests em março de 2020

3.1.1 Etapa de Extração dos Dados

Na etapa de extração dos dados, representado com a cor azul na Figura 3.1, o objetivo é obter uma amostra representativa dos repositórios do GitHub. Inicialmente, foram identificadas as 15 linguagens mais utilizadas no GitHub no mês de março de 2020. Para tanto, foi utilizada a ferramenta de data warehouse tool BigQuery ² para se extrair as linguagens mais utilizadas no GitHub no período - a contabilização foi baseada no número de pull requests constantes no GitHub Archive. A Figura 3.2 apresenta essas linguagens e a respectiva quantidade de pull requests realizados. Posteriormente, foi realizada a busca dos 100 repositórios mais populares de cada uma dessas linguagens por meio da GitHub API utilizando como medida de popularidade o número de estrelas recebidas pelos repositórios. Ressalta-se que essa mesma métrica já foi usada como representante de popularidade em outros estudos no âmbito de MSR [15, 70, 42]. Essa primeira etapa obteve um conjunto de 1.500 repositórios.

Em seguida foram obtidos os dados dos 1.500 repositórios do dataset inicial. Contudo, diante dos desafios relacionados ao grande número de dados a serem obtidos, foram utilizadas duas abordagens para a extração dos dados necessários para o cálculo dos indicadores: (i) os dados referentes a commits foram obtidos por meio do clone dos repositórios de forma local, valendo-se dos utilitários do git para a extração dos dados; e (ii) a utili-

²<https://cloud.google.com/bigquery>

zação da API do GitHub para obtenção dos dados referentes a pull requests. O processo de extração dos dados dos repositórios foi automatizado utilizando scripts escritos na linguagem de programação Python, que estão disponíveis no pacote disponibilizado para fins de reprodução do processo. Para acesso à API do GitHub foi utilizada a biblioteca PyGitHub³. Os dados extraídos cobrem dois períodos distintos: antes da pandemia da COVID-19 e durante a pandemia. Foi considerado como início da pandemia a data em que a OMS declarou o surto de COVID-19 como uma pandemia, ou seja, em 11 de março de 2020. Contudo, foram excluídos dados referentes aos 30 dias anteriores e posteriores a essa data, uma vez que diferentes regiões adotaram medidas restritivas em diferentes épocas e isso poderia enviesar as análises. Nesse contexto, o período observado durante a pandemia inicia em 10 de abril de 2020 e termina em 10 de abril de 2021 - o que dá um total de 12 meses. Por fim, para propósitos de comparação, também foram obtidos os dados dos 12 meses anteriores à declaração da pandemia (excluído o período de transição), dessa forma, o período anterior a pandemia a ser observado inicia em 9 de fevereiro de 2019 e termina em 9 de Fevereiro de 2020.

Adicionalmente, foram excluídos repositórios com baixo nível de atividades no período e repositórios que não remetem ao armazenamento de um projeto de software. Para tanto, foram excluídos os repositórios que estão abaixo do primeiro quartil na distribuição de número de commits, número de pull requests e número de contribuidores. O motivo para filtrar os repositórios dessa maneira é garantir que os projetos analisados remetam a um software e utilizem o paradigma de pull request. Conseqüentemente, os repositórios com um nível consistente de atividades têm um maior potencial para demonstrar possíveis impactos nos indicadores em função da pandemia. Dessa forma, o dataset final possui **155** repositórios. A Tabela 3.1 apresenta a estatística descritiva dos repositórios observados.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|----------------|-------|---------|---------|---------|---------|-----------|
| Linhas de Cod. | 8.300 | 88.820 | 169.096 | 333.757 | 347.393 | 3.272.254 |
| Stars | 2.413 | 12.523 | 19.900 | 30.954 | 37.030 | 318.358 |
| Forks | 288 | 2.403 | 4.193 | 8.159 | 8.886 | 83.669 |
| Contribuidores | 145 | 350 | 633 | 1.060 | 1.167 | 13.238 |

Tabela 3.1: Estatística descritiva dos repositórios observados

3.1.2 Etapa de Modelagem

Nessa etapa, representada pela cor verde na Figura 3.1, os dados foram modelados, uma vez que foram extraídos de forma não estruturada e havia a necessidade de se combinar dados de mais de uma origem: commits extraídos de forma local e pull requests extraídos

³<https://github.com/PyGithub/PyGithub>

por meio da API. Além disso, a análise temporal a ser empregada necessita que algumas variáveis expressem a passagem do tempo e torne possível a comparação entre os dois períodos (antes e depois do início da pandemia). Nesse sentido, foram criadas três variáveis artificiais que modelam a passagem do tempo e possibilitam a identificação de uma descontinuidade no nível dos indicadores com o advento da pandemia:

- **time**: variável numérica que compreende os meses do início até o fim do período observado (antes e durante a pandemia), dessa forma, a variável `time` tem 24 valores possíveis, representando os meses observados.
- **time_after_intervention**: variável numérica e tem valor 0 nos doze meses anteriores à pandemia e os valores de 1 até 12 nos meses posteriores ao início da pandemia.
- **intervention**: variável do tipo lógica e possui o valor `False` para o período anterior ao início da pandemia e `True` no período que representa os meses em que a pandemia ocorria.

Em seguida, foi realizada a identificação dos core developers através do Algoritmo 1, que implementa a heurística utilizada no estudo realizado por Ferreira et al. [1]. Essa heurística consiste em identificar desenvolvedores que foram autores de 80% dos commits dos repositórios no período de 12 meses, excluindo desenvolvedores que contribuíram com um valor menor que 5% do total dos commits do período. Em seguida foi feita uma segunda identificação dos core developers, com o auxílio de uma ferramenta automatizada [47].

A partir da identificação dos core developers, com as duas abordagens citadas, foi realizada a implementação do cálculo da taxa de Turnover, usando as fórmulas especificadas na Seção 2.3. O Algoritmo 2 demonstra o cálculo da taxa de turnover levando em consideração dois períodos: o mês a ser calculado o turnover ($Período_i$) e o mês imediatamente anterior ($Período_{i-1}$). Na linha 6 é calculada a taxa de Core Developers Leavers (`CDL_rate`) que representa a taxa de desenvolvedores que deixaram de ser Core Developers no período. Já na linha 7 é realizado o cálculo de pessoas que passaram a ser Core Developers no período. A linha 8 do Algoritmo 2 remete à taxa de Core Developer Turnover (`CDT_rate`) e armazena o valor da taxa de turnover no repositório no período i . O cálculo desses indicadores foi executado em todos os meses analisados e listados na Tabela 3.2.

Adicionalmente, foram identificados os core developers utilizando uma outra abordagem: o Truck Factor. Nesse caso, uma ferramenta automatizada foi utilizada para extrair os desenvolvedores [47]. Para que os resultados também fosse agrupados em períodos de

Algorithm 1 Cálculo de Core Developers

```
1: function CALCULACD(commits) ▷ commits do repositório em determinado período
2:   num_commits = count(commits)
3:   sum commits grouped by author
4:   sort sum commits desc
5:   core_developers = {}
6:   total = 0
7:   for each author a in commits sorted do
8:     if sum commits of a ≤ 0.05 then
9:       break
10:    end if
11:    if total ≥ num_commits * 0.8 then
12:      break
13:    end if
14:    total = total + sum commits of a
15:    core_developers.insert(a)
16:    return core_developers
17:  end for
18: end function
```

30 dias e assim, analisar o turnover mês a mês, foi realizado o checkout dos repositórios locais no início de cada mês e, então, comparada a lista de desenvolvedores em relação ao mês anterior. Em seguida o turnover foi realizado também seguindo as fórmulas de [1]. Embora a abordagem de contabilização de commits possa identificar desenvolvedores que também são identificados pela abordagem do TruckFactor, inclusive estudos anteriores demonstram esse fato [39], é importante verificar se as duas abordagens remetem a resultados diferentes e assim ter uma análise mais ampla.

Algorithm 2 Cálculo do Turnover de Core Developers

```
1: function CALCULATURNOVER(CommitsPeriodi-1, CommitsPeriodi)
2:   SETA = CalculaCD(CommitsPeriodi-1)
3:   SETB = CalculaCD(CommitsPeriodi)
4:   CDL = each dev ∈ SETA and ∉ SETB           ▷ Core Developers Leavers
5:   CDN = each dev ∉ SETA and ∈ SETB           ▷ Core Developers NewComers
6:   CDL_rate = (length(CDL)/length(SETA))*100
7:   CDN_rate = (length(CDN)/length(SETB))*100
8:   CDT_rate = (CDL/((length(SETA)+length(SETAB))/2))*100
9:   return core_developers           ▷ Retorna a taxa de turnover do Períodoi
10: end function
```

Por fim, nessa etapa também foram implementados o cálculo das contribuições casuais e da taxa de ingresso de novos participantes nos projetos, utilizando os mecanismos mencionados no Capítulo 2.

3.1.3 Etapa de Síntese

Nessa etapa, representada pela cor amarela na Figura 3.1, os dados relacionados aos indicadores do nível de atividade dos repositórios foram agrupados considerando janelas de 30 dias e identificados pelas variáveis temporais já mencionadas. Os dados referentes a cada período são apresentados na Tabela 3.2. Ressalta-se que o agrupamento utilizando a unidade de tempo mensal já foi usado por outros estudos no escopo da Engenharia de Software [32, 33].

Em relação ao número de commits e número de pull requests, foram somadas as ocorrências dentro de cada unidade de tempo por repositório. Já em relação aos indicadores commit churn e pull request latency foram utilizadas as medianas dos valores em cada mês dos dois períodos observados por repositório. A escolha da mediana como unidade de tendência central ocorreu porque os dados não seguem uma distribuição normal, ou seja, os dados possuem caudas pesadas ou deformações em relação à sua distribuição [71]. Além disso, durante a execução dos modelos de regressão, os dados foram transformados para a escala logarítmica, também por não possuírem distribuição normal. Essa recomendação feita por Hemmati et al. [22], também foi utilizada em outros trabalhos que utilizaram modelos de regressão baseados em variáveis que representam o nível de atividade em repositórios de software [33, 32].

A etapa de síntese também compreende a classificação e agrupamento de dados [22]. Dessa forma, nessa etapa também foram classificados os sentimentos dos 2.661.886 comentários de pull requests dos repositórios do dataset realizados durante os 24 meses observados. Inicialmente, os comentários foram extraídos utilizando a API do GitHub e armazenados em um arquivo do tipo CSV; além do comentário, foram extraídos metadados relacionados a cada registro identificado. Para a classificação da polaridade do sentimento (entre negativo, neutro ou positivo), foi utilizada a ferramenta SentiSD [55]. Para tanto, um modelo foi treinado utilizando uma base de dados pública [58] com anotações de mais de 7.000 documentos do GitHub. Essa abordagem possui um desempenho superior a outras ferramentas do tipo para a classificação dos sentimentos [58]. Em seguida foi realizada a classificação dos 2.661.886 comentários; o resultado da classificação (negativo, neutro ou positivo) foi então inserido no arquivo CSV com os dados dos comentários. Por fim, os comentários foram agrupados em bases mensais, baseado em sua data de criação e a proporção da polaridade dos sentimentos foi computado.

3.1.4 Etapa de Análise

Na etapa final, representada pela cor vermelha na Figura 3.1, foram realizadas análises utilizando os dados obtidos. Essas análises tem como objetivo entender como a crise

| Data Início | Data Fim | Time | Intervention | Time After Intervention |
|-------------|------------|------|--------------|-------------------------|
| 15/02/2019 | 17/03/2019 | 1 | False | 0 |
| 17/03/2019 | 16/04/2019 | 2 | False | 0 |
| 16/04/2019 | 16/05/2019 | 3 | False | 0 |
| 16/05/2019 | 15/06/2019 | 4 | False | 0 |
| 15/06/2019 | 15/07/2019 | 5 | False | 0 |
| 15/07/2019 | 14/08/2019 | 6 | False | 0 |
| 14/08/2019 | 13/09/2019 | 7 | False | 0 |
| 13/09/2019 | 13/10/2019 | 8 | False | 0 |
| 13/10/2019 | 12/11/2019 | 9 | False | 0 |
| 12/11/2019 | 12/12/2019 | 10 | False | 0 |
| 12/12/2019 | 11/01/2020 | 11 | False | 0 |
| 11/01/2020 | 10/02/2020 | 12 | False | 0 |
| 10/04/2020 | 10/05/2020 | 13 | True | 1 |
| 10/05/2020 | 09/06/2020 | 14 | True | 2 |
| 09/06/2020 | 09/07/2020 | 15 | True | 3 |
| 09/07/2020 | 08/08/2020 | 16 | True | 4 |
| 08/08/2020 | 07/09/2020 | 17 | True | 5 |
| 07/09/2020 | 07/10/2020 | 18 | True | 6 |
| 07/10/2020 | 06/11/2020 | 19 | True | 7 |
| 06/11/2020 | 06/12/2020 | 20 | True | 8 |
| 06/12/2020 | 05/01/2021 | 21 | True | 9 |
| 05/01/2021 | 04/02/2021 | 22 | True | 10 |
| 04/02/2021 | 06/03/2021 | 23 | True | 11 |
| 06/03/2021 | 05/04/2021 | 24 | True | 12 |

Tabela 3.2: Períodos observados e as variáveis temporais inseridas.

global causada pela disseminação do COVID-19 interferiu nas contribuições realizadas na plataforma GitHub. Como os dados não obedecem uma distribuição normal, foram realizados testes não-paramétricos para testar possíveis diferenças na distribuição dos indicadores nos dois períodos: antes e depois da pandemia. Além disso, para se ter uma dimensão do impacto, foi utilizada para a mensuração do tamanho do efeito a abordagem do Cliff's Delta [72].

Com o intuito de reforçar a análise dos dados, foi aplicada na análise a mesma abordagem utilizada por Wessel et al. [33] e foram criados modelos do tipo *Regression Discontinuity Design* (RDD). Um modelo RDD tem o potencial de avaliar o efeito causal de uma intervenção, verificando se houve uma descontinuidade de determinado indicador no momento de uma intervenção e também ao longo do tempo. Nesse caso, foi considerado o início da pandemia como o momento da intervenção e, dessa forma, avaliado se isso criou uma descontinuidade dos indicadores (variáveis dependentes) observados durante o início da pandemia. Também como o realizado no estudo realizado por Wessel et al. [33], a

passagem do tempo foi modelada adicionando três variáveis que agregam a característica de série temporal ao modelo: `time`, `time_after_intervention` e `intervention`. A variável `time` é numérica e compreende os meses do início até o fim do período observado (antes e durante pandemia), dessa forma, a variável `time` tem 24 valores possíveis, representando os meses observados. Já a variável `time_after_intervention` também é numérica e tem valor 0 nos doze meses anteriores à pandemia e os valores de 1 até 12 nos meses posteriores ao início da pandemia. Por fim, a variável `intervention` é do tipo lógica e possui o valor `False` no período anterior ao início da pandemia e `True` no período que representa os meses em que a pandemia ocorria.

Essas variáveis de séries temporais tem o potencial de mensurar o impacto sofrido pelos indicadores no início da pandemia e nos meses seguintes. Com o auxílio dessas variáveis, também é possível avaliar se já existia uma tendência de aumento ou queda antes do início da pandemia. Essas características, são importantes para se ter uma análise mais detalhada dos reais impactos da pandemia no nível de atividade dos repositórios. O modelo de regressão pode ser especificado como:

$$y_i = \alpha + \beta \times time_i + \gamma \times intervention_i + \delta \times time_after_intervention_i + \epsilon_i$$

Dessa forma, a inclinação da linha de regressão antes da pandemia tem o valor de β , e depois da intervenção (no caso o início da pandemia) $\beta + \delta$. O tamanho do efeito do tratamento é a diferença entre os dois valores no momento da intervenção e é igual a γ [32].

Além das variáveis temporais, compõem o modelo as variáveis de controle agregadas mensalmente. Em cada modelo foram utilizadas apenas como variáveis explicativas, aquelas que possuem fator de inflação da variância (VIF), menor do que 5 e, assim, evitar a multicolinearidade.

- **log(commit_churn)**: variável numérica que indica a mediana das somas das linhas inseridas e deletadas no repositório. Essa variável é importante pois mensura o esforço produzido em cada commit.
- **log(num_commits)**: quantidade mensal de commits do repositório. Por meio do número de commits espera-se mensurar o trabalho dos desenvolvedores que efetivamente foi adicionado ao repositório.
- **log(num_pr)**: quantidade mensal de pull requests do repositório. Diferente dos commits, os pull requests podem capturar um nível de atividade diferente, pois pode ser criado por qualquer desenvolvedor, inclusive os novos contribuidores.

- **log(latency_pr)**: a mediana em horas do tempo entre a criação e fechamento do pull request no mês observado. Por normalmente envolver mais de um ator e depender de um processo de revisão, espera-se capturar o tempo gasto com o processo de revisão.
- **time_after_creation**: variável numérica, que expressa em meses o tempo desde a criação do repositório. Importante para verificar se repositórios mais antigos possuem um nível de atividade diferente.
- **log(num_contributors)**: variável numérica que representa o número de contribuidores que o repositório possui. Nesse caso, essa variável pode demonstrar se a quantidade de contribuidores afeta diretamente o nível de atividade.
- **log(num_newcomers)**: variável numérica que representa o número de novos desenvolvedores (que nunca haviam contribuído com o repositório, desde a sua criação). A variável foi adicionada aos modelos para mensurar o efeito que novos contribuidores tem nas variáveis dependentes.

Para que o modelo possa expressar a variabilidade causada pela aleatoriedade e não explicada pelas variáveis independentes, foi utilizada uma abordagem mixed-effects [73]. Dessa forma, além das variáveis de fixed-effects apresentadas acima, foram adicionadas duas variáveis que representam random-effects:

- **language**: a linguagem de programação principal utilizada no repositório.
- **repository**: variável contendo o nome de cada repositório do dataset.

Com a adição das variáveis aleatórias, é possível verificar quanto da variabilidade do nível de atividade é causada pela simples mudança de linguagem de programação e de repositório para repositório [33], [32]. Além do estudo de Wessel et al. [33], já citado anteriormente, outros estudos no escopo da Engenharia de Software utilizaram modelos RDD com mixed-effects para verificar o impacto de intervenções em repositórios do GitHub [32, 74].

3.2 Survey

Desenho do Survey

Foram realizados dois surveys: o survey 1 investiga os efeitos da pandemia na produtividade e no bem-estar das equipes de desenvolvimento de software e, dessa forma, responde a QP.4. Já o survey 2, tem como público alvo os core developers, e visa verificar os efeitos

da pandemia sob o ponto de vista desse tipo de desenvolver e também subsidiar a triangulação dos dados do esforço de mineração. Ambos surveys foram disponibilizados como questionários na plataforma online Google Forms ⁴. O survey 1 possui 26 questões (24 questões fechadas e 2 questões abertas), organizadas em seções com temas demográficos, com o impacto da pandemia na vida pessoal e o impacto da pandemia na produtividade relacionada aos projetos de código aberto. O survey foi construído de forma a capturar como fatores humanos e organizacionais influenciaram os membros dos times de projetos de software de código aberto em função das medidas de isolamento. Dessa forma, o questionário levanta questões relacionadas ao perfil do participante como: país, o papel no projeto, experiência, etc. Também foram relatadas questões relacionadas à percepção da produtividade nos projetos em que o participante contribui e uma autoavaliação da produtividade, antes e depois das medidas de isolamento social decorrentes da pandemia. Os participantes também foram encorajados a compartilhar suas percepções em duas questões abertas: na primeira para que fossem feitos comentários sobre como a pandemia impactou a produtividade do próprio desenvolvedor e na segunda, para que fossem feitos comentários sobre como os repositórios que o desenvolvedor contribui foram impactados. O questionário foi baseado em um estudo anterior [6], com adaptações para o contexto de projetos OSS. Já o survey 2 visa identificar a percepção dos core developers em relação aos projetos e também os efeitos da pandemia nesse grupo de desenvolvedores. Além das questões demográficas, foram adicionadas questões específicas sobre a percepção dos core developers em relação aos níveis de atividade dos repositórios; eles foram questionados sobre variações, após o início da pandemia, no números de commits, pull requests, issues, comentários, turnover e ingresso de novos contribuidores. O survey 2 possui no total de 39 questões, sendo 35 questões fechadas e 4 questões abertas.

Audiência do Survey

No esforço de mineração, foram extraídos todos os e-mails disponíveis a partir dos logs de commits de todos os repositórios analisados. Nesse contexto, o esforço de mineração teve como uma das saídas **74.098** e-mails distintos. Considerando a dificuldade em se enviar uma quantidade tão grande de e-mails foram filtrados apenas e-mail com o domínio do Gmail. A lista final de e-mails aptos a receberem o survey 1 foi de **25.983** contribuidores. O questionário ficou disponível de 19 de Setembro de 2020 a 04 de Novembro de 2020. No total, 137 participantes responderam ao survey 1 e as questões podem ser acessadas na url <https://forms.gle/5MMmZCSC7bYVMgfWA>. Após a obtenção dos contribuidores, foi realizada a identificação dos core developers utilizando a mesma heurística de [1]. Essa heurística identificou 1274 core developers durante os 24 meses analisados; contudo,

⁴<https://www.google.com/intl/pt-BR/forms/about/>

desse total apenas 485 possuíam um e-mail disponível no perfil. Dessa forma, 485 core developers foram convidados a responder ao questionário, sendo obtida 57 respostas, o que dá uma taxa de resposta de 11.7%. O questionário do survey 2 está disponível em <https://forms.gle/aBeaRAMqnKE4VyoSA> e ficou disponível de 22/07/2021 a 19/08/2021.

Análise do Survey

Em relação ao survey 1, inicialmente foi executado o tratamento dos dados com a remoção das respostas de dois participantes que apresentaram respostas fora dos padrões: um deles preencheu valores negativos quando perguntado sobre a quantidade de crianças em casa e outro participante preencheu frases sem sentido no campo destinado a uma pergunta discursiva. Todas as respostas do survey 2 foram consideradas válidas.

Adicionalmente, para preservar a anonimidade dos dados foi apagado o campo referente ao carimbo de tempo e separadas as questões abertas dos dois surveys em arquivos a parte. Em seguida foi calculada a diferença entre a auto-produtividade e a produtividade do time relatada pelos participantes antes e depois da pandemia. Por fim, foram realizadas análises baseadas em estatística descritiva e o uso de boxplots para análise das respostas.

3.3 Síntese do Capítulo

Neste capítulo foram apresentados detalhes do estabelecimento do esforço de mineração, que foi dividido em quatro etapas (extração, modelagem, síntese e análise). De forma mais específica, foram descritos os detalhes para modelagem dos indicadores relacionados aos core developers e desenvolvedores casuais. Além disso, foi descrita a abordagem utilizada nos modelos de regressão RDD utilizado para se identificar os efeitos da pandemia. Por fim, foram descritos detalhes do design, execução e análise dos dois surveys.

Capítulo 4

Resultados

Este capítulo apresenta um detalhamento em relação aos resultados dos estudos quantitativo e qualitativo, obtidos por meio do esforço de mineração com 155 repositórios, do survey 1 com 135 desenvolvedores, do survey 2 com os core developers dos projetos e da análise de sentimentos expressos em artefatos gerados durante a comunicação dos contribuidores dos repositórios.

4.1 QP.1 - Qual o impacto da pandemia nas atividades de desenvolvimento nos repositórios de projetos de software de código aberto?

Para responder essa questão de pesquisa, inicialmente cinco indicadores foram analisados como representantes do nível de atividades do repositório: números de commits, commit churn, número de pull requests, a latência de pull request e a frequência dos comentários. A Figura 4.1 sintetiza os resultados relacionados às atividades de commits e pull requests, mostrando a mediana dos indicadores coletados no período que compreende 12 meses *anteriores* do início das medidas de isolamento (meses com índices de -12 até -1) e nos 12 meses *posteriores* ao início das medidas de isolamento (meses com índices de 1 até 12). Em seguida são demonstrados as variações nas medidas de tendência central e realizados testes estatísticos (utilizando o teste de Mann-Whitney-Wilcoxon [75]) para verificar a diferença na distribuição dos indicadores considerando os dois períodos e também o tamanho do efeito (utilizando a abordagem do Cliff's delta [76]). Adicionalmente, foi construído um modelo de regressão utilizando como variável dependente cada um dos indicadores estudados e reportados os resultados quando o indicador possuir significância estatística (considerando o p-value < 0.05) entre as distribuições dos dois períodos. Por fim, são reportados os resultados do survey 2, com a percepção dos core developers sobre

o nível de atividade dos repositórios, incluindo os indicadores analisados no processo de MSR.

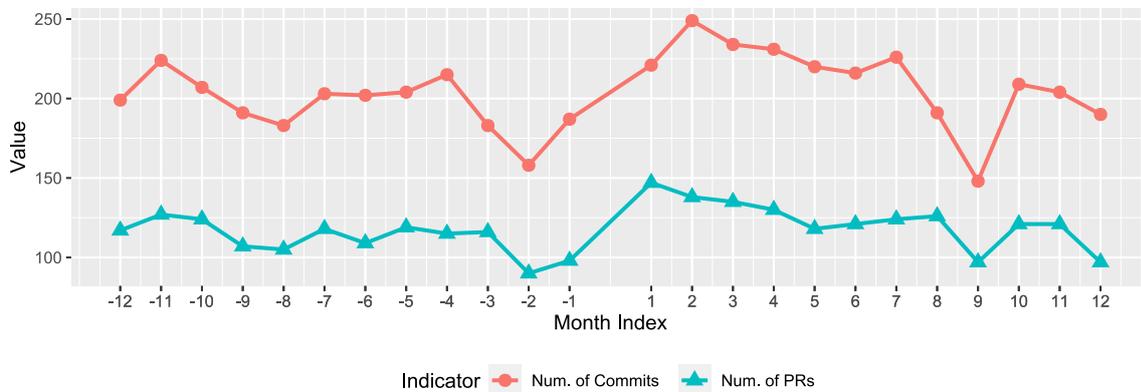


Figura 4.1: A mediana dos indicadores de atividade dos repositórios agrupados de forma mensal

Número de commits

De forma geral, considerando os 155 repositórios de forma conjunta, a mediana do número de commits mensal antes da pandemia era de 196 commits. Depois do início da pandemia a mediana subiu para 209 commits. Apesar da pequena diferença da mediana dos dois períodos, houve diferença significativa na distribuição das frequências, utilizando o teste Mann-Whitney-Wilcoxon ($p\text{-value} = 0.0001574$), mas o effect size foi insignificante (-0.01327234 , usando a abordagem do Cliff's delta).

Contudo, quando analisados os repositórios de forma individual, há uma diferença significativa ($p\text{-value} < 0.05$), em relação ao número de commits, em 76 (49.03%) dos 155 repositórios analisados. Desses 76 repositórios, 44 deles apresentaram um effect-size negativo, com uma mediana do Cliff's delta de -0.7361111 (grande), caracterizando uma diminuição na frequência de commits; por outro lado, 32 repositórios tiveram um effect-size positivo, com uma mediana do Cliff's delta de 0.6284722 (grande), indicando um aumento no número de commits.

A Tabela 4.1 contém os resultados do modelo de regressão que tem o número de commits como variável dependente. Pode-se observar que houve uma descontinuidade positiva significativa ($p\text{-value} < 0.05$) no nível de commits logo após o início da pandemia, dado que coeficiente da variável *intervention* possui um valor positivo e estatisticamente significativo. Ao longo dos meses, uma pequena tendência de aumento se manteve, uma vez que a soma dos coeficientes das variáveis *time* e *Time_after_intervention* resulta em um valor positivo, mas em uma magnitude menor do que a variável *intervention*. Esse mesmo padrão pode ser observado no gráfico da Figura 4.2 e na Figura 4.1: embora

exista uma inclinação positiva no mês cujo índice é 1 (e que corresponde ao início da intervenção), seguindo de um pico no mês de índice 2, é possível notar que ao longo dos meses a mediana de commits retornou a um nível similar ao período anterior à pandemia.

Em relação aos outros coeficientes do modelo de regressão da Tabela 4.1, é possível notar que, além da intervenção, a variável $\log(size)$, que representa a quantidade de linhas do projeto, é a que mais tem efeito sobre o número de commits; esse resultado é esperado, uma vez que sistemas maiores, em termos de linhas de código, tendem a ter um maior nível de atividade. Sob o ponto de vista do R^2 , o modelo obteve um R_m^2 de 0.38, mas grande parte da variabilidade pode ser atribuída às variáveis aleatórias uma vez que o R_c^2 foi 0.83.

Para entender como repositórios com características diferentes (em termos de número de contribuidores e tamanho) se comportaram ao longo da pandemia, os repositórios foram agrupados de acordo com essas características e analisadas as diferenças em relação aos indicadores. A Tabela 4.2 demonstra a distribuição dos 155 repositórios, agrupados por número de participantes no projeto. Já a Tabela 4.3 demonstra a distribuição dos repositórios em relação ao tamanho do projeto (utilizando como medida o número de linhas de código). Nesse último caso, percebe-se uma maior concentração de repositórios (25) com efeito positivo no grupo formado pelos que têm mais linhas de código e uma mediana do Cliff's delta maior (0.549, o que é considerado grande) em relação aos grupos que tiveram efeito positivo; essa característica está alinhada com o modelo de regressão, no qual a variável $size$ é responsável por explicar a maior parte da variação dos dados no modelo de regressão que tem o número de commits como variável dependente (veja Tabela 4.1).

Observação 1. Foi observado um aumento expressivo no número de commits no início da pandemia, mas houve uma estabilização ao longo dos meses. Quando considerados apenas os repositórios com diferença estatística significativa (76), o número de repositórios com diminuição (44/76) é maior que o número de repositórios com aumento no efeito (32/76) na frequência de commits.

Commit Churn

A fim de se ter uma visão sobre o impacto na quantidade de trabalho realizado, foi calculada a mediana de *commit churn* em cada mês dos repositórios observados. Inicialmente, o commit churn de todos os commits foi calculado somando-se o total de linhas excluídas mais as linhas inseridas em cada commit [32]. Quando analisado esse indicador em todos os repositórios, foi obtida uma mediana de commit churn de 17 nos dois períodos (antes e após o advento da pandemia). Essa mesma estabilidade pode ser observada na Figura

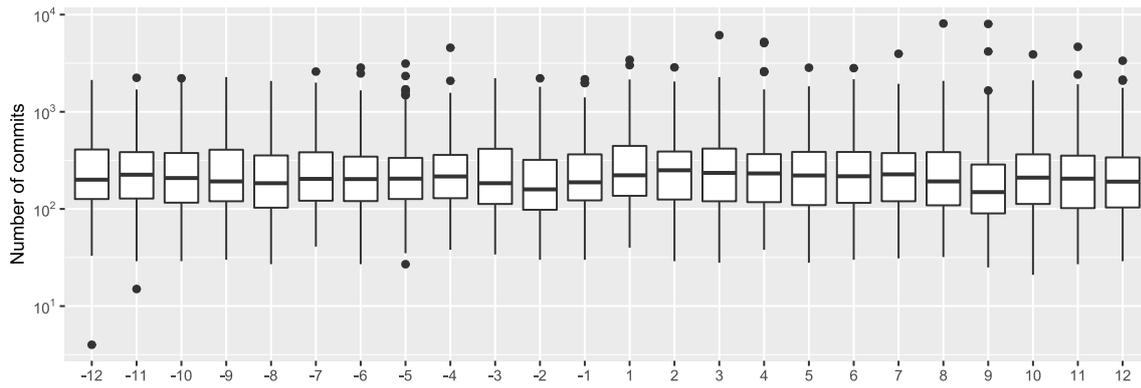


Figura 4.2: Série temporal dos commits antes e depois da pandemia.

| Variable | Coefficients | Sum Sq. |
|-------------------------|--------------|---------|
| time | 0.005* | 0.582 |
| intervention | 0.163*** | 8.302 |
| time_after_intervention | 0.008** | 1.121 |
| log(commit_churn) | 0.067*** | 0.957 |
| log(CDT_rate) | 0.019*** | 1.676 |
| log(size) | 0.355*** | 11.426 |
| time_after_creation | 0.003 | 0.375 |
| log(num_contributors) | 0.048 | 3.351 |
| log(num_newcomers) | 0.015*** | 63.481 |

Note: *p<0.1; **p<0.05; ***p<0.01

Tabela 4.1: O efeito da pandemia no número de commits. A variável dependente é $\log(\text{número de commits})$. $R_m^2 = 0.38$. $R_c^2 = 0.83$.

| Contribuidores | Efeito Positivo | | Efeito Negativo | |
|----------------|-----------------|--------------------------|-----------------|--------------------------|
| | Projetos | Mediana Cliff's Delta | Projetos | Mediana Cliff's Delta |
| 145-347 | 19 | 0.444 | 19 | -0.562 |
| 348-632 | 18 | 0.319 | 20 | -0.615 |
| 633-1.171 | 15 | 0.431 | 25 | -0.458 |
| 1.172-13.238 | 23 | 0.354 | 16 | -0.625 |
| Total | 75 | 0.3819 | 80 | -0.5729 |

Tabela 4.2: Repositórios que apresentaram diferença significativa em relação à distribuição do número de commits antes e durante a pandemia, divididos em quartis de acordo com o número de contribuidores.

4.3, uma vez que não existe grandes variações ao longo do tempo. Contudo, foi encontrada uma diferença significativa na distribuição do indicador (p-value = 0.003398 e Cliff

| LOC | Efeito Positivo | | Efeito Negativo | |
|-------------------|-----------------|--------------------------|-----------------|--------------------------|
| | Projetos | Mediana Cliff's Delta | Projetos | Mediana Cliff's Delta |
| 8.300-88.816 | 13 | 0.292 | 25 | -0.451 |
| 88.816-169.096 | 24 | 0.351 | 15 | -0.458 |
| 169.096-352.012 | 13 | 0.417 | 26 | -0.569 |
| 352.012-3.272.254 | 25 | 0.549 | 14 | -0.733 |
| Total | 75 | 0.3819 | 80 | -0.5729 |

Tabela 4.3: Repositórios que apresentaram diferença significativa em relação à distribuição do número de commits antes e durante a pandemia, divididos em quartis de acordo com o número de linhas de código.

delta = -0.033268) - considerando o conjunto de repositórios analisados. Ao analisar os repositórios de forma individual, dos 155 repositórios analisados, 38 (24.51%) demonstraram uma diferença significativa em relação ao commit churn depois da intervenção. De forma específica, 13 desses repositórios tiveram um effect-size grande e positivo (com uma mediana de Cliff's delta de 0.5486); e 25 repositórios tiveram um effect-size negativo (com uma mediana do Cliff's delta de -0.625).

Observação 2. Tanto a mediana quanto o gráfico não indicam alterações em relação ao commit churn antes e durante a pandemia. Considerando apenas os repositórios com diferença estatística significativa (38), houve uma maior quantidade com efeito negativo (25/38) do que positivo (13/38).

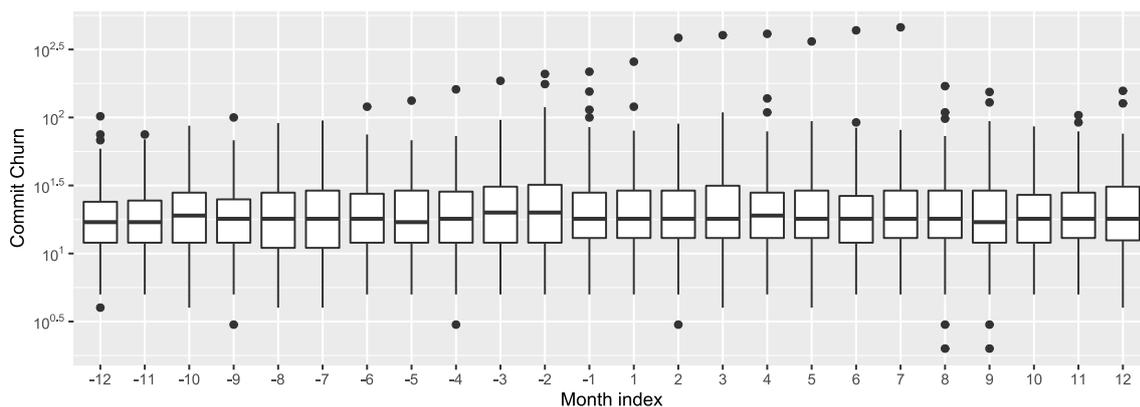


Figura 4.3: Série temporal do commit churn antes e depois da pandemia.

Número de Pull Requests

O modelo de desenvolvimento baseado em pull requests é amplamente usado em plataformas de hospedagem de código, como o GitHub [77]. O processo completo compreende uma série de passos e possui ao menos dois atores: o contribuidor e o integrador [37]; dessa forma, esse processo pode fornecer evidências sobre o grau de atividade nos repositórios de uma forma mais ampla. Também foram empregadas as medianas do número de pull requests agrupadas por mês. E considerando a atividade de todos os repositórios do dataset de forma conjunta, a mediana antes da pandemia foi de 111 pull requests e depois do início da pandemia a mediana foi de 124, um aumento de 11.7%.

Houve uma diferença significativa quando comparadas as distribuição dos pull requests antes e depois da pandemia: o p-value é de 9.257e-16 e o Cliff's delta é insignificante (-0.05889). Quando analisados de forma individual foi observado que 76 repositórios (49.03%) tiveram uma diferença estatística significativa, mas nem todos repositórios tiveram a mesma direção: 50 deles tiveram um effect-size negativo (com a mediana do Cliff's de -0.75) e 26 repositórios tiveram um effect-size positivo (com uma mediana do Cliff's delta de 0.6041). Os coeficientes do modelo que tem o número de pull requests como variável dependente (veja a Tabela 4.5) indicam que houve uma descontinuidade no nível de pull requests logo após o início da pandemia, mantendo-se todas as variáveis constantes. Essa descontinuidade, tem um impacto positivo no número de pull requests e explica muito da variabilidade dos dados. Contudo, pode-se observar que já havia uma pequena tendência de crescimento antes da pandemia (variável *time* positiva e estatisticamente significativa) e esse padrão se manteve ao longo dos meses do período de pandemia, uma vez que a soma das variáveis *time* e *time_after_intervention* têm como resultado um valor positivo. Esse mesmo padrão pode ser observado no gráfico da Figura 4.1 e na Figura 4.4.

Em relação ao outros coeficientes do modelo de regressão disponíveis na Tabela 4.5 é possível notar que os coeficientes das variáveis $\log(\text{num_contributors})$ e $\log(\text{size})$ explicam muito da variabilidade no número de pull requests; esse resultado é esperado, uma vez que sistemas maiores, em termos de linhas de código e participantes, tendem a ter um maior nível de pull requests. Esse mesmo comportamento pode ser verificado na Tabela 4.4, o conjunto de repositórios com mais contribuidores possui o maior número de repositórios com diferença estatística positiva. Contudo, a variável $\log(\text{num_newcomers})$ também é responsável por grande parte da variabilidade do modelo, indicando que novos contribuintes podem ter tido um efeito positivo sobre o número de pull requests. Sob o ponto de vista do R^2 , o modelo obteve um R_m^2 de 0.44, mas grande parte da variabilidade pode ser atribuída às variáveis aleatórias uma vez que o R_c^2 foi 0.86.

Observação 3. Houve um aumento expressivo no número de pull requests no início da pandemia. Dos repositórios que apresentaram diferença significativa (76 no total), uma maior parte dos repositórios apresentou diminuição (50/76) e uma parcela menor apresentou aumento (26/76).

| Contribuidores | Efeito Positivo | | Efeito Negativo | |
|----------------|-----------------|--------------------------|-----------------|--------------------------|
| | Projetos | Mediana Cliff's Delta | Projetos | Mediana Cliff's Delta |
| 145-347 | 16 | 0.347 | 22 | -0.549 |
| 348-632 | 15 | 0.396 | 23 | -0.569 |
| 633-1.171 | 12 | 0.267 | 28 | -0.493 |
| 1.172-13.238 | 17 | 0.417 | 22 | -0.458 |
| Total | 60 | 0.3680 | 95 | -0.5208 |

Tabela 4.4: Repositórios que apresentaram diferença significativa em relação à distribuição do número de PRs antes e durante a pandemia, divididos em quartis de acordo com o **número de contribuidores**.

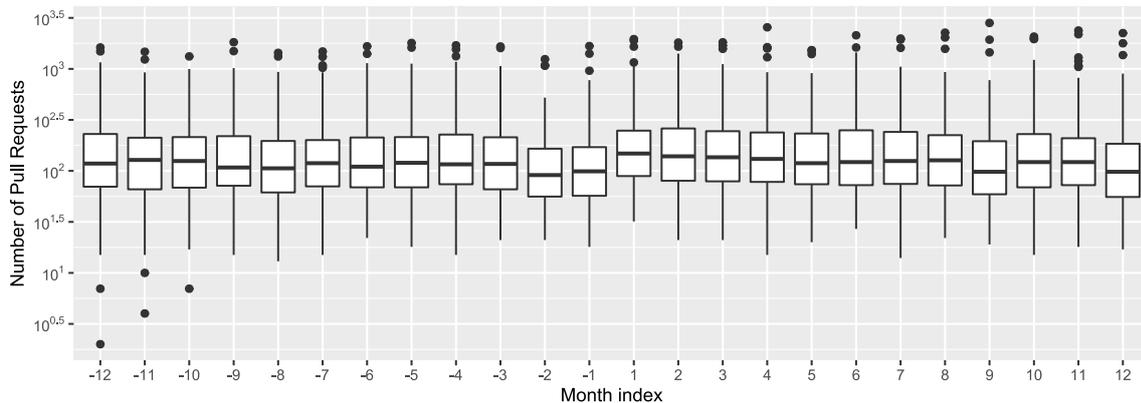


Figura 4.4: Série temporal do número de pull requests antes e depois da pandemia.

Latência de Pull Requests

Também foi observada a latência dos pull requests usando a mediana do tempo (em horas) entre a criação e o fechamento dos pull requests. De forma mais ampla, comparando a mediana de todos os repositórios de forma conjunta, no período anterior à pandemia a mediana era de 21.62 horas e depois do início da pandemia a mediana foi de 20.84 horas, uma diminuição de 3.6%. Nesse contexto, não foi encontrada diferença significativa entre a distribuição da latência ($p\text{-value} = 0.3754$), comparando os a latência dos pull requests dos 155 repositórios de forma conjunta. E isso vai ao encontro do gráfico da Figura 4.5, uma

vez que não há grandes variações ao longo dos meses. Contudo, ao analisar os repositórios individualmente, foram encontradas diferenças significativas em 45 repositórios (29.03%). Desses repositórios, 23 apresentaram um effect-size positivo, com um Cliff's delta de 0.6805 (grande); de forma contrastante, 22 repositórios apresentaram um effect-size negativo, com uma mediana do Cliff's delta de -0.7083 (grande).

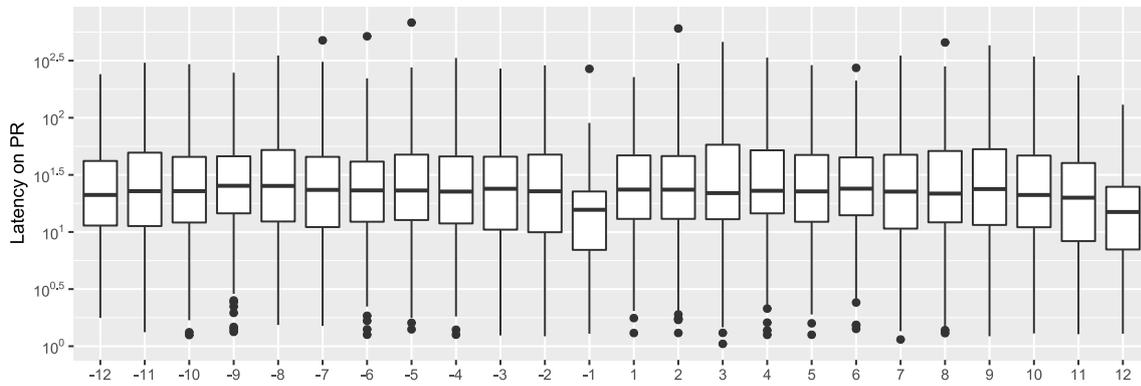


Figura 4.5: Série temporal da latência de pull requests antes e depois da pandemia.

| Variable | Coefficients | Sum Sq. |
|-------------------------|--------------|---------|
| time | 0.013*** | 0.374 |
| intervention | 0.360*** | 24.387 |
| time_after_intervention | 0.012*** | 2.036 |
| log(latency_pr) | 0.174*** | 44.526 |
| log(CDT_rate) | 0.015*** | 0.967 |
| log(size) | 0.250*** | 8.030 |
| time_after_creation | 0.002 | 0.024 |
| log(num_contributors) | 0.254*** | 7.056 |
| log(num_newcomers) | 0.283*** | 59.560 |

Note: *p<0.1; **p<0.05; ***p<0.01

Tabela 4.5: O efeito da pandemia no número de pull requests. A variável dependente é o $\log(\text{número de pull requests})$. $R_m^2 = 0.44$. $R_c^2 = 0.86$.

Observação 4. Não foi encontrada diferença estatística significativa na distribuição da latência de pull requests antes e durante a pandemia. Além disso, há uma pequena diferença (3.6%) entre a mediana observada antes e durante a pandemia.

Comentários

Uma vez que os pull requests são submetidos, eles passam por um processo de revisão. Durante esse processo, dois tipos de comentários podem ser feitos: discussões, que remetem

ao conteúdo geral do pull request e revisões de código, que referenciam a seções específicas do código, destacando a diferença entre o código atual e o código proposto [18]. Dessa forma, para entender os impactos da pandemia na comunicação entre os atores no processo de revisão de pull requests, foram extraídos os comentários dos repositórios observados durante os 12 meses antes e após o início da pandemia. De uma forma geral, a mediana mensal de comentários antes da pandemia era de 232.5, e, após o início da pandemia, a mediana aumentou para 249.5, um aumento de 7.3%. Ainda comparando os dois períodos, foi encontrada uma diferença significativa na distribuição desse tipo de comentário nos dois períodos (p-value = 0.02087), mas o tamanho do efeito foi insignificante (cliff's delta = -0.03042057). A análise individual mostra que dos 155 repositórios analisados, 60 apresentaram diferença significativa; sendo 25 com effect size positivo, e uma mediana do cliff's delta de 0.6527778; e 35 com effect size negativa, nesse caso, a mediana do cliff's delta foi de -0.7083333. Nos dois casos o tamanho do efeito é considerado grande, mas apresentando direções opostas. A Figura 4.6 demonstra a distribuição dos comentários a cada mês e a figura 4.7 a mediana dos comentários a cada mês. É possível notar nos dois gráficos que houve um aumento na mediana no início da pandemia, seguindo com uma diminuição na frequência dos comentários ao longo dos meses.

Observação 5. O número de comentários teve um aumento expressivo durante os primeiros meses da pandemia. Dos repositórios que apresentaram diferença estatística significativa (60), houve mais repositórios com diminuição (35/60) do que com aumento (25/60) na quantidade de comentários.

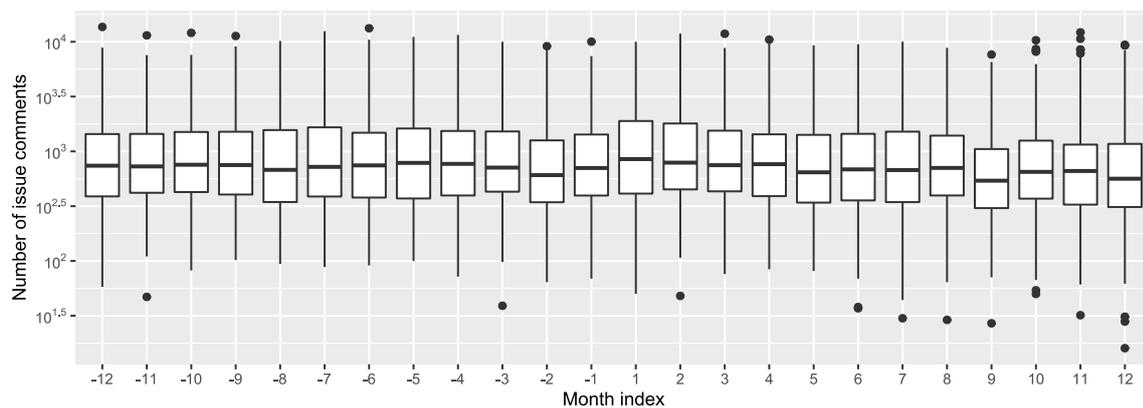


Figura 4.6: Série temporal do número de review comments antes e depois da pandemia.

Percepção dos Core Developers

A fim de se ter uma perspectiva a partir dos membros dos projetos sobre os indicadores analisados até aqui, o Survey 2 teve como audiência 57 core developers identificados

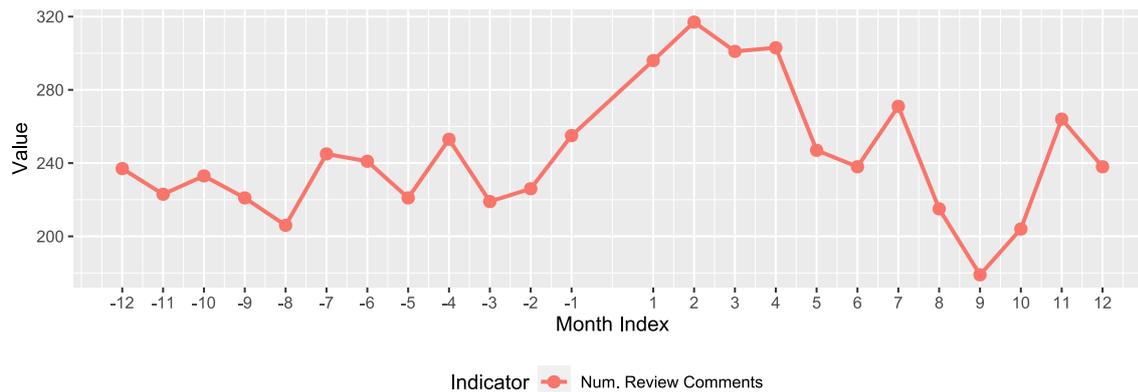


Figura 4.7: A mediana dos comentários dos repositórios agrupados de forma mensal

a partir dos 155 repositórios estudados. Os participantes foram questionados sobre sua percepção de produtividade, sobre os nível de atividades gerais e sobre alterações específicas nos indicadores. Os core developers possuem uma visão única sobre as práticas de desenvolvimento porque possuem engajamento e conhecimento em relação aos projetos; a saída deles inclusive configura uma ameaça à continuidade do projeto, devido à perda de conhecimento e experiência [16]. Dessa forma, os core developers podem ajudar a validar os achados do esforço de mineração e a lançar luz sobre outros assuntos.

Em relação à produtividade, quando os core developers foram questionados sobre sua própria produtividade por meio da questão "Please, indicate how the COVID-19 pandemic has changed your productivity", 31,6% deles reportaram um aumento, 38,6% indicaram que a produtividade se manteve a mesma e 29,8% informou que a própria produtividade diminuiu em função da pandemia. Esses resultados vão ao encontro da percepção que os core developers tiveram sobre os seus colegas contribuidores, pois quando perguntados diretamente o quanto concordavam com a afirmação "Do you agree that the COVID-19 outbreak negatively impacted on the productivity of your colleagues to contribute to the de código aberto projects you contribute to?", 48% deles disseram não concordar com a afirmação, 19% se mantiveram neutros e 33% disseram concordar com a afirmação. Dessa forma, aproximadamente 30% dos participantes acreditam que a própria produtividade diminuiu durante a pandemia e 33% deles também acreditam que produtividade dos colegas diminuiu.

Em relação a percepção do nível de atividade de uma maneira geral nos projetos em que acompanham, a maioria dos core developers também indicou que não houve mudanças, pois quando perguntados "How would you compare the contributions for open source projects before and during the pandemics?", 47,4% afirmou que o nível de contribuições manteve-se o mesmo. Ainda, 31,6% indicou uma diminuição no nível de atividades e 21,1% indicou um aumento. Especificamente em relação aos indicadores, os core developers fo-

ram encorajados a indicar suas percepções sobre variações nos níveis dos indicadores que também foram minerados no MSR, por meio da questão: "Please indicate your perceptions regarding changes in level of activities on projects you use to contribute, after the start of COVID-19 pandemic." A síntese das respostas pode ser observado na Figura 4.8. Depreende-se da figura, que a maioria os participantes acredita que não houve alterações em nenhum dos indicadores, mas dos que pensam ter existido mudanças, a maioria considera que houve um aumento na latência dos PRs, mas também um aumento no nível de de atividades relacionadas à commits, pull requests e comentários. Curiosamente, a mesma quantidade de participantes (22%) indicou aumento e diminuição no tamanho dos commits.

Observação 6: A percepção dos participantes do survey 2 sobre os efeitos da pandemia na própria produtividade é contrastante: 31.6% deles reportaram um aumento, 38,6% indicaram estabilidade e 29,8% informaram que a própria produtividade diminuiu.

Observação 7: Entre os participantes que reportaram mudanças no nível de atividade dos repositórios, uma maior parte acredita em diminuição (31.6%); em contrapartida 21.1% acreditam em aumento. Contudo, especificamente em relação aos indicadores, a maioria reportou um aumento em todos eles.

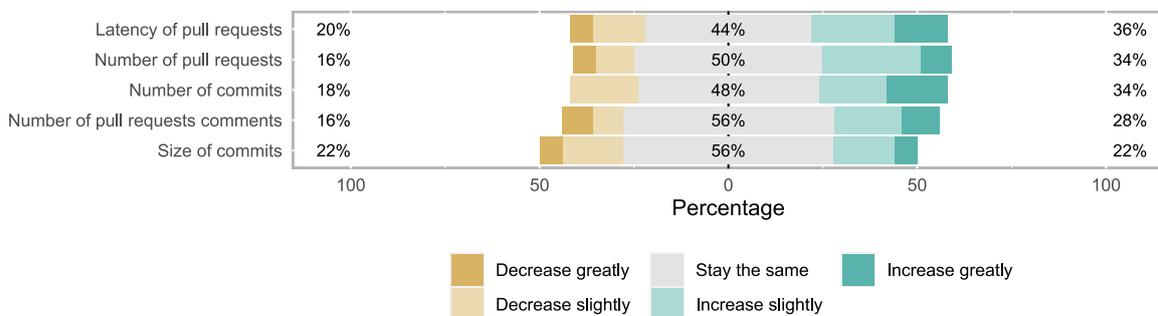


Figura 4.8: A percepção dos core developers sobre mudanças nos indicadores de atividade dos repositórios causados pela pandemia.

4.2 QP.2 - Qual o impacto da pandemia no ingresso de novos contribuidores?

O ingresso contínuo e o engajamento de novos contribuidores com o desenvolvimento do projeto é crucial para o sucesso de projetos de código aberto [17]. E embora já existissem

várias barreiras para a entrada e retenção de novos membros [78] [79] mesmo em tempos normais, é importante investigar se o ingresso desses desenvolvedores foi afetado durante a pandemia, Dessa forma, para responder a QP.2 foi investigado o ingresso de novos contribuidores e também o engajamento nos sistemas observados ao longo do período de 12 meses.

Em relação à entrada de novos contribuidores, no período anterior à pandemia a média mensal de ingresso nos repositórios era de 16.28 contribuidores por mês; e nos 12 meses após o início da pandemia a média foi de 14.29, uma redução de 12.2%. Ao observar o impacto nos repositórios de forma individual, houve uma maior quantidade com redução na média após o início da pandemia: 63% dos repositórios tiveram uma redução na média, enquanto 34% tiveram um aumento e 3% não tiveram alterações. Embora exista uma significância estatística ($p\text{-value} = 3.77e-08$) em relação à média de entrada de novos contribuidores antes e durante a pandemia usando o teste de Mann-Whitney-Wilcoxon, o efeito da diferença é insignificante (0.0625), usando a abordagem do Cliff's Delta.

Quando analisados o número de novos contribuidores em cada mês, é possível observar um leve aumento na média de entrada de participantes nos dois primeiros meses durante a pandemia, mas ainda menor do que meses observados antes da pandemia. Além disso, essa leve tendência de alta se inverteu ao longo dos meses com uma tendência de queda no ingresso de novos contribuidores ao longo do período de pandemia, como visto na Figura 4.9. Com o intuito de verificar se essa queda já estava acontecendo em períodos anteriores ou se realmente ocorreu exclusivamente durante a pandemia, as médias anuais de ingresso também foram computadas e agrupadas, observado o ano durante a pandemia e também três anos anteriores. Pode-se perceber uma diminuição nesse indicador no ano que compreende a pandemia (índice 1), conforme Figura 4.10. A média de ingresso no ano observado durante a pandemia foi de 150.74 novos contribuidores por repositório, nesse caso são considerados apenas os 132 repositórios criados a partir de 2017. Esse valor é menor do que o observado nos três anos anteriores (índices -1, -2 e -3 na Figura 4.10).

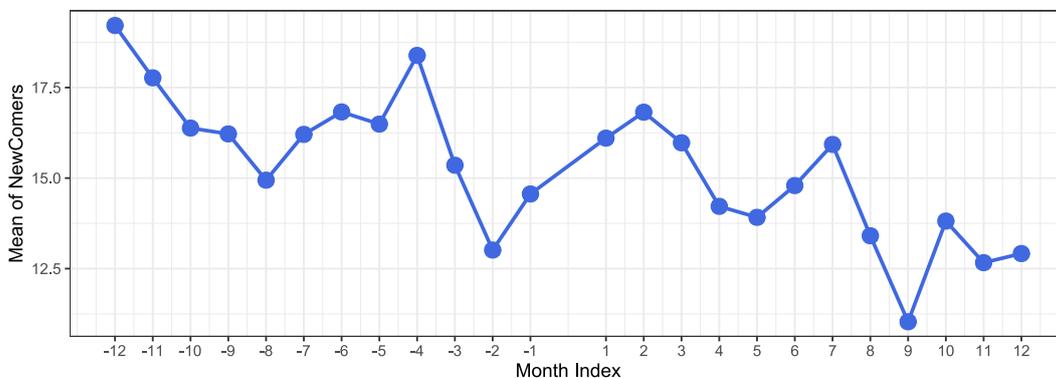


Figura 4.9: A média de novos contribuidores ao longo dos meses.

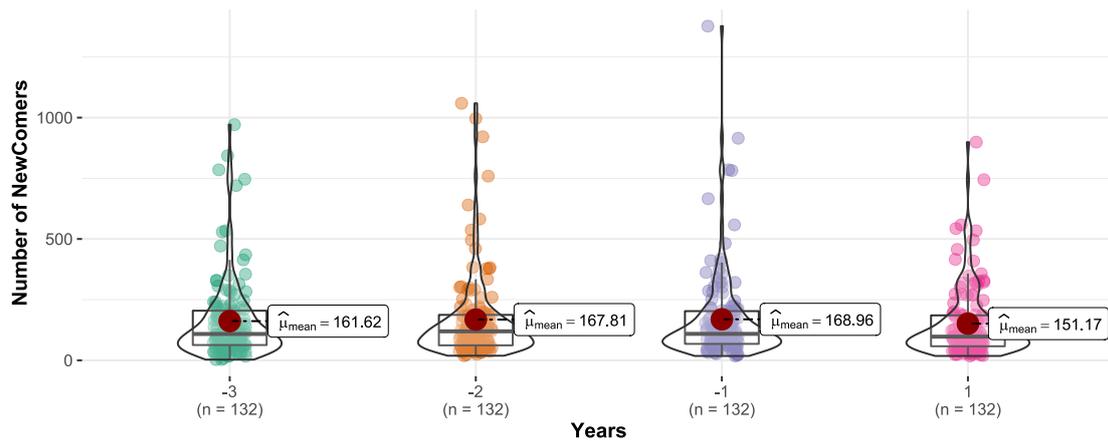


Figura 4.10: Ingresso de novos contribuidores ao longo dos anos.

A falta de conhecimento sobre a linguagem de programação de um projeto é uma das dificuldades enfrentadas no ingresso de novos contribuidores nos projetos [80]. Dessa forma, a falta de domínio de uma linguagem específica pode afetar o ingresso de novos contribuidores. Nesse contexto, a Tabela 4.6 mostra as linguagens de programação mais usadas nos 155 repositórios analisados, e uma comparação entre o ingresso de novos contribuidores antes e durante a pandemia. É possível notar que os projetos escritos em Javascript e Go tiveram as maiores diminuições no ingresso de contribuidores (-32.59% e -16.24%, respectivamente). A única linguagem de programação da lista que teve um leve aumento foi a linguagem C# (1.38%).

Em relação ao engajamento, a Figura 4.12 mostra o número total de novos contribuidores a cada mês e também o número de desenvolvedores que fizeram a sua contribuição em um determinado mês e que não contribuíram novamente no mesmo projeto; os índices com valores negativos representam o período anterior à pandemia, enquanto os índices 1 e 2 representam os dois primeiros meses observados durante a pandemia. Foram inseridos apenas os dois primeiros meses do segundo período para que exista uma janela de 12 meses sem atividades e, somente assim, considerar que o desenvolvedor deixou de contribuir; além disso, esses dois primeiros meses foram os que apresentaram a maior alta de ingresso durante a pandemia. A Figura 4.11 também remete ao mesmo período; mas nesse caso é demonstrada a taxa de contribuidores que fizeram uma única contribuição, em relação ao total de novos contribuidores. É possível notar que não houve grandes variações no percentual dos valores no primeiro e segundo mês durante a pandemia: 56.60% e 57.34% dos desenvolvedores respectivamente abandonaram o projeto após a sua primeira e única contribuição; para fins de comparação, a média dos 12 meses anteriores é de 54.78%. Essas constatações indicam que não houve um aumento expressivo no engajamento entre os novos contribuidores no início da pandemia.

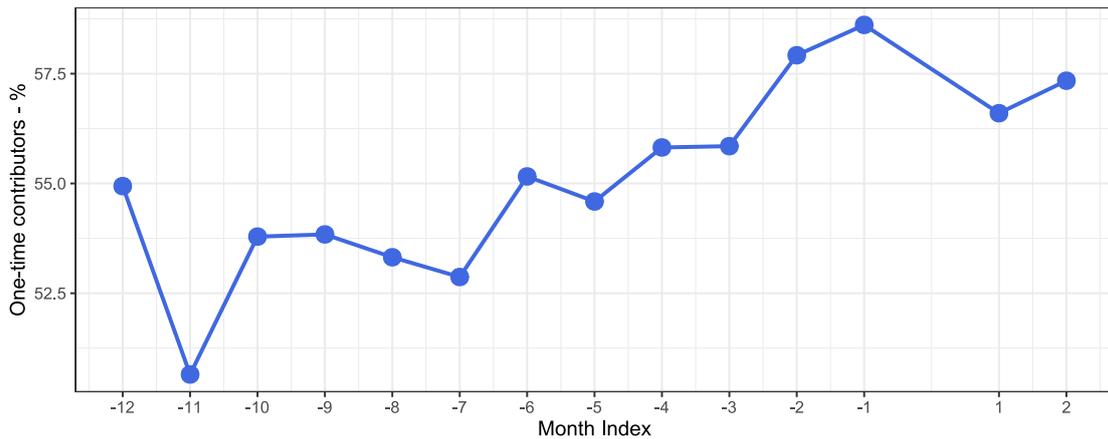


Figura 4.11: Percentual de desenvolvedores que só fizeram um commit em relação a todos os novos contribuidores, ao longo dos meses.

Observação 8: Houve uma diminuição de 12.2% na média mensal de ingresso de novos contribuidores após o início da pandemia. Além disso, 63% dos repositórios apresentaram redução no ingresso no mesmo período. O histórico também indica que o período da pandemia foi o que teve maior redução na série histórica de 4 anos.

| Ling. progr. | Repositórios | Total Antes | Total Durante | Variação |
|--------------|--------------|-------------|---------------|----------|
| Go | 22 | 3213 | 2691 | -16.24% |
| C++ | 21 | 3700 | 3477 | -6.02% |
| TypeScript | 21 | 5677 | 5223 | -7.99% |
| JavaScript | 19 | 5338 | 3598 | -32.59% |
| C# | 13 | 1159 | 1175 | 1,38% |
| Python | 12 | 3728 | 3440 | -7,72% |
| Java | 11 | 1426 | 1346 | -5,61% |
| Ruby | 10 | 2767 | 2536 | -8,34% |

Tabela 4.6: Diferença no ingresso de novos contribuidores depois do início da pandemia em relação à linguagem de programação usada nos repositórios

Observação 9: Os projetos escritos nas linguagens de programação Javascript e Go apresentaram as maiores reduções no ingresso de novos contribuidores. A redução foi de -32.59% e -16.24%, respectivamente.

Para dimensionar as mudanças nas contribuições casuais após o início da pandemia, foram computadas a taxa desse tipo de contribuição em relação ao total de commits no período anterior e posterior à declaração da pandemia. Nesse estudo, contribuidores casuais são aqueles desenvolvedores que realizaram apenas um commit durante a existência do repositório [43]. Dessa forma, a análise de contribuição casual remete aos desenvolvedores

que contribuíram uma única vez e não voltaram a contribuir novamente durante o período observado. Inicialmente, foi calculada a média de contribuições de todos os repositórios nos dois períodos, e em seguida o percentual dessas contribuições que são consideradas casuais. O resultado sugere que a pandemia não impactou diretamente a taxa de contribuições casuais, dado que a média de contribuições casuais antes da pandemia era de 5,57% e depois da pandemia, a média foi de 5,59%, uma diferença de apenas 0,003%.

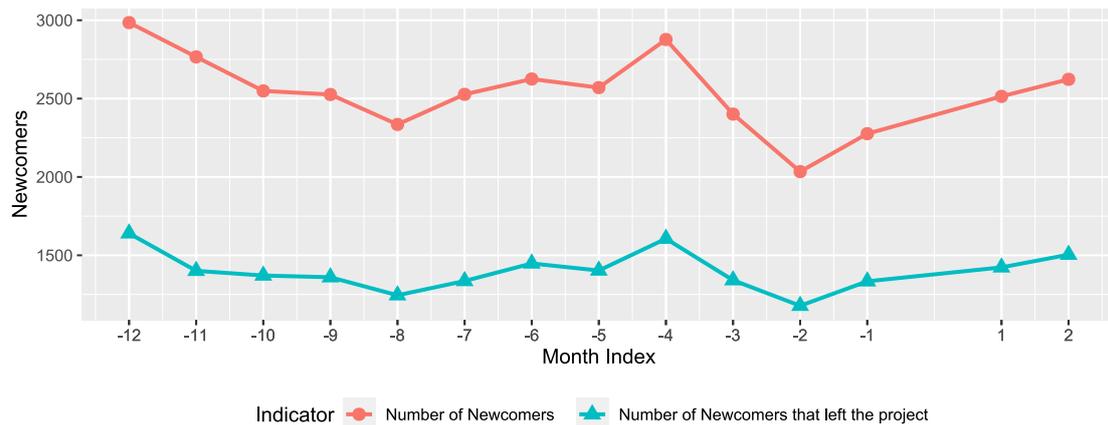


Figura 4.12: Número de Newcomers vs. Newcomers que deixaram de contribuir com o projeto após a primeira contribuição.

4.3 QP.3 - Qual o impacto da pandemia no turnover de core developers?

Para responder a questão de pesquisa QP.3, inicialmente foram computados o turnover mensal dos core developers utilizando duas abordagens: a identificação dos core developers com a contabilização dos commits [1] e a identificação utilizando a abordagem do TruckFactor [47].

Na primeira abordagem, no período anterior à pandemia a média mensal de turnover foi de 5.21%, enquanto que nos doze meses posteriores ao início da pandemia essa taxa caiu para 4.64%, uma diminuição de aproximadamente 11%. Além disso, dos 155 repositórios analisados, 79 (50.97%) tiveram uma diminuição na taxa de turnover, 60 (38,71%) tiveram um aumento no turnover de core developers e 16 (10.32%) não tiveram mudanças na taxa de turnover, comparando os dois períodos.

Em relação aos repositórios que tiveram aumento do turnover, 11 deles (7%) tiveram algum turnover de core developers apenas em meses posteriores ao início da pandemia; entre esses repositórios estão alguns bem populares como `rust-lang/rust`, `twbs/bootstrap`, e `spring-projects/spring-framework`.

De forma mais granular, quando observado o turnover mês a mês, é possível notar um pico de quase 11% na taxa de turnover (Figure 4.13) logo após o início da pandemia (month index = 1). De fato, o mês representado pelo índice 1 foi o mês que apresentou a maior quantidade de repositório com turnover de core developers: 48 (30.97%) apresentaram algum turnover nesse mês específico. Além disso, a média de repositórios com turnover de core developers a cada mês - considerando todos os 24 meses observados (12 antes e 12 após a intervenção) - foi de 27.58. Dessa forma, a quantidade de repositórios que apresentaram turnover no primeiro mês após o advento da pandemia é aproximadamente 74% mais alto que a média de todos os meses observados.

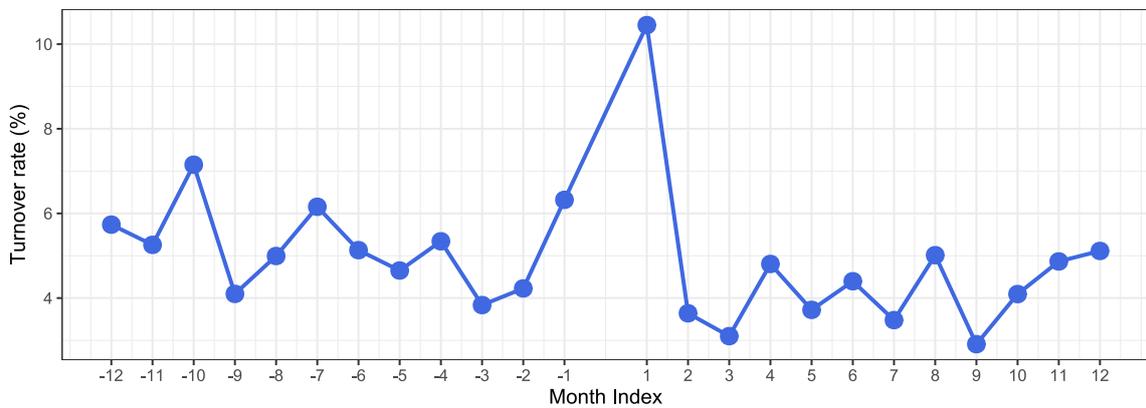


Figura 4.13: Média da taxa de turnover de core developers, utilizando a contabilização de commits [1] %

A taxa de turnover é calculada levando em conta o número de core developers que abandonaram o projeto (Core Developer Leaver), mas pode ser atenuada caso novos contribuidores se tornem core developers no período (Core Developer Newcomer) [1]. Dessa forma, é importante verificar a tendência desses dois indicadores nos períodos observados. A média mensal de Core Developers Leavers antes do início da pandemia era de 4.66% e depois do início da pandemia a média foi de 4,11%, uma diminuição de 11%. Já a média de Core Developers Newcomers foi de 5.09% antes da pandemia e de 5.08% durante a pandemia, uma diminuição de apenas 0,001%. A taxa de *Core Developers Newcomers* e *Core Developers Leavers* também tiveram as maiores mudanças no primeiro mês do início da pandemia (month index = 1): 11.13% e 9.23% respectivamente. A fim de se ter a a visão dos próprios core developers sobre o assunto, durante o survey 2 eles foram perguntados sobre o quanto concordavam com a afirmação "There are developers that used to be part of core developers before the pandemic, but stopped being core developer after the start of the pandemic." e também com a afirmação "Some developers became core developer only after the start of the pandemic."; curiosamente, 28% afirmaram concordar fortemente ou concordar com as duas afirmações.

A análise do turnover realizada até aqui não faz distinção entre desenvolvedores que deixaram de ser core developer em um mês e que voltaram a ser core developer em um mês posterior e pessoas que não voltaram a ser core developers no projeto ao longo dos meses. Portanto, caso um desenvolvedor deixe de ser core developer no mês de índice 1 e volte a ser core developer no mês de índice 3, o desenvolvedor será considerado um Core Developer Leaver no mês 1 e um Core Developer NewComer no mês 3. Para mensurar a quantidade de pessoas que deixaram de ser core developer e que não voltaram a esse posto, foi verificada a taxa de Core Developer Leavers no início da pandemia e que não retornaram a essa posição ao longo dos meses. Nesse sentido, dos 582 core developers existentes no primeiro mês da pandemia, 50 (8,5%) deles não voltaram a ser core developers nos 11 outros meses analisados posteriormente, considerando os 155 repositórios analisados. Para fins de comparação, essa taxa foi de 5.1% no mesmo período do ano anterior à pandemia.

Outro fator importante, são os core developers que abandonaram o projeto (e não apenas deixaram de ser core developers); para investigar esse ponto foi calculado o número de core developers que deixou de fazer contribuições por 12 meses. Esse mesmo período foi utilizado em [39] e, embora eles possam voltar a contribuir no futuro, a probabilidade não é considerada grande [39]. Nesse sentido, a Figura 4.14 mostra o número de core developers que saíram do projeto em cada mês (sendo que o mês contabilizado, indica o último commit realizado pelo desenvolvedor). É possível notar que oito core developers deixaram de contribuir com os projetos no primeiro mês observado dentro da pandemia; em contrapartida a média anterior (considerando os 12 meses anteriores à pandemia) foi de 10.25. Essa abordagem levou em conta apenas os commits, não sendo consideradas outras atividades, como por exemplo criação de issues, mentoria, gerenciamento de projetos, etc.

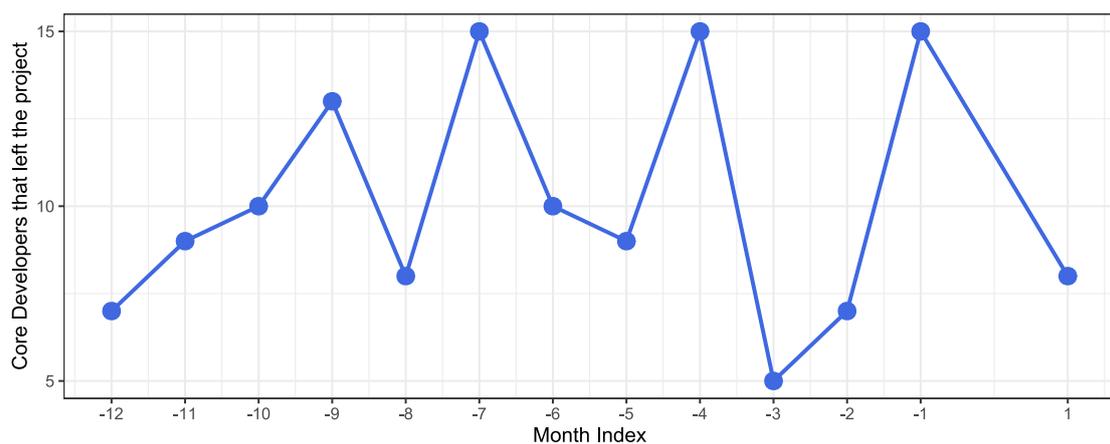


Figura 4.14: Número de core developers que deixaram de contribuir com o projeto

Para se ter uma visão histórica em relação a esse indicador, foi computado o turnover anual de core developers considerando os 4 anos anteriores à pandemia. A Figura 4.15

contém gráficos do tipo violino ao longo dos anos; o gráfico com a legenda 1 representa o ano em que houve pandemia e os outros gráficos (-1 a -3) representam os anos anteriores. Com efeito, foram considerados apenas os repositórios que foram criados a partir de 2017, o que dá um total de 132 projetos. É possível notar que no ano em que ocorreu a pandemia houve um menor nível de turnover do que nos anos anteriores, sendo que a média de turnover dos repositórios nesse período (índice 1 na Figura 4.15) foi de 32.17%.

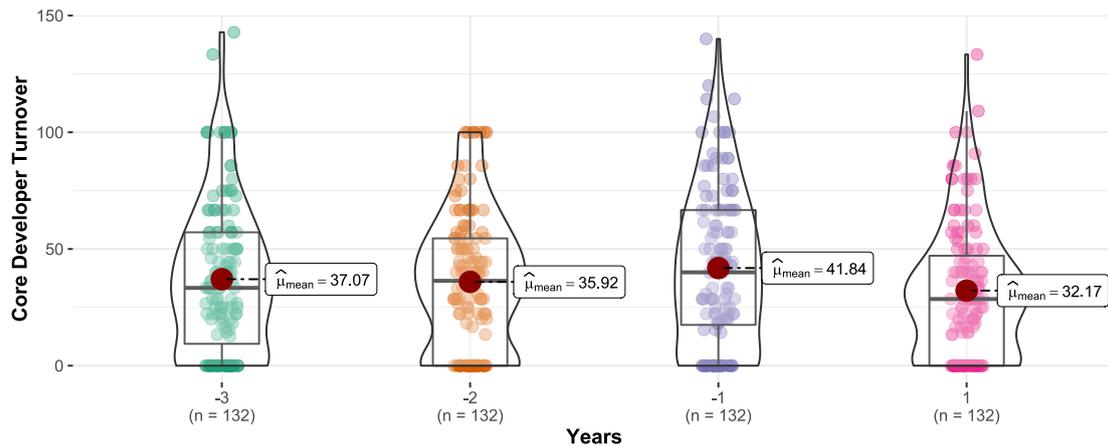


Figura 4.15: Comparação do Turnover de Core Developers ao longo dos anos, utilizando a contabilização de commits [1] %

Também foi analisado o turnover em relação ao nível de atividades dos repositórios, foi computado o turnover do ano durante a pandemia e agrupados os repositórios de acordo com a quantidade de commits. Para tanto, os grupos na Figura 4.16 remetem aos quartis do número de commits dos repositórios durante o período da pandemia. É possível notar na Figura 4.16 que repositórios com um maior nível de commits possuem um menor nível de turnover de core developers. O gráfico cuja legenda é Q4, e que representa os repositórios mais ativos em termos de commits (4.274 a 55.333 commits durante o ano em que houve pandemia), possui uma média de commits menor do que os outros grupos (Q1, Q2 e Q3), e que representam um menor nível de contribuições. Essa distribuição não foi observada no ano anterior à pandemia, nesse caso houve uma maior taxa de turnover de core developers justamente em repositórios menos ativos (Q1 = 37.69% e Q4 = 44.25%).

A segunda abordagem, que identifica os core developers com o algoritmo do Truck Factor [47], demonstra que o período anterior à pandemia teve uma média de turnover de core developers de 7.88% e que a média no período durante a pandemia foi de 7.90%, considerando todos os repositórios analisados. Apesar da pequena diferença na média dos dois períodos, é possível perceber dois picos durante os meses de índice 1 e 4 (ambos dentro do período posterior ao início da pandemia), conforme Figura 4.17. Esses meses

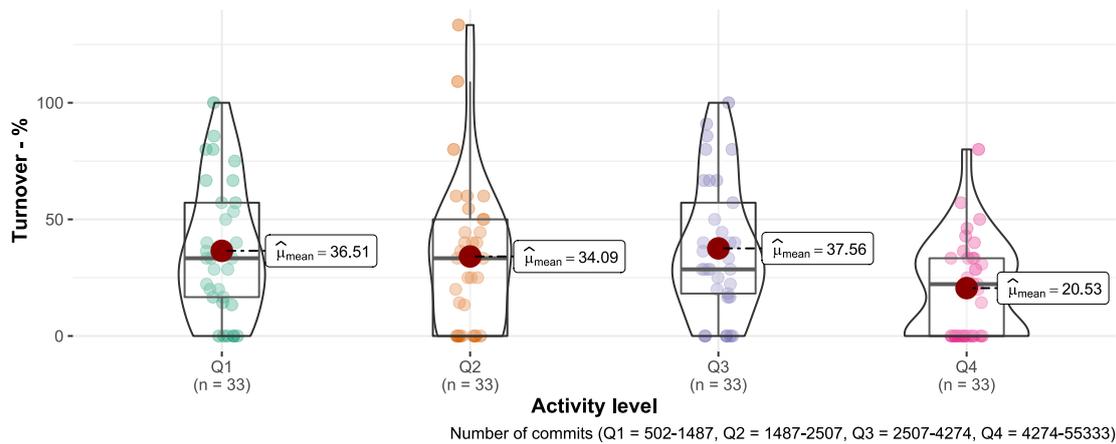


Figura 4.16: Comparação do turnover de core developers em relação a repositórios com diferentes níveis de atividade, utilizando a contabilização de commits [1] %

tiveram, respectivamente, 11.13% e 11.37% de turnover médio, considerando todos os repositórios.

A abordagem do Truck Factor, também identificou alguns repositórios experimentando algum turnover apenas no período posterior ao início da pandemia. De forma mais específica, 14 repositórios se encaixam nessa situação.

Já em relação aos desenvolvedores que deixaram de ser core developer, a média da taxa de Leavers foi de 7.43% antes e de 7.55% após o início da pandemia, uma pequena diferença próxima a 1%. Além disso, a média da taxa de desenvolvedores que se tornaram core developers diminuiu de 8.59% para 8.52% por mês. Também foi computada a taxa de core developers após deixarem esse grupo no primeiro mês e que não retornaram a esse posto nos meses seguintes: dos 889 Core Developers, 54 (6%) deles não voltaram a ser um core developer ao longo da pandemia; de forma paralela, no período anterior à pandemia, 618 desenvolvedores eram considerados core developers no primeiro mês observado antes da pandemia (índice -12), sendo que 32 (5.1%) deles não voltaram a esse posto nos 11 meses seguintes.

Por fim, a primeira abordagem (com contabilização de commits) apresentou maiores variações na média; enquanto a abordagem do Truck Factor, teve médias mais equilibradas nos dois períodos. Contudo, a primeira abordagem teve apenas um pico na média das taxas de turnover logo no início da pandemia (Figura 4.13). Enquanto a segunda abordagem apresentou esse mesmo pico e um outro no 4º mês ainda durante o período de pandemia (Figura 4.17).

Observação 10: Ao usar as duas abordagens (contabilização de commits e TruckFactor), foi observado um pico de turnover mensal no 1º mês da pandemia, porém não

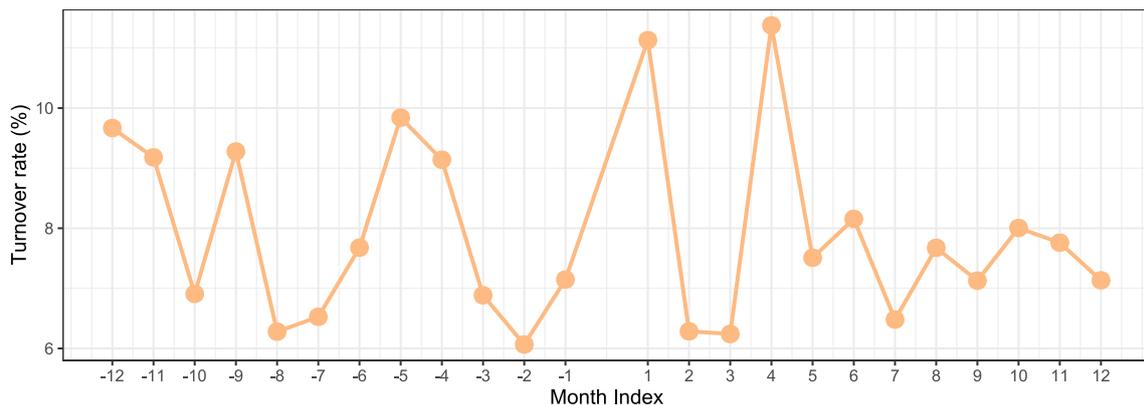


Figura 4.17: Média da taxa de turnover de core developers, utilizando o algoritmo de TruckFactor %

foi percebido aumento considerável no desengajamento dos core developers durante esse pico: do total de 582 core developers, apenas 8 abandonaram os projetos.

4.4 QP.4 - Qual a influência da pandemia no estado emocional e bem estar dos desenvolvedores de OSS?

Para responder a questão de pesquisa QP.4, inicialmente foi executado um survey com contribuidores de projetos de OSS. No total o survey 1 teve **135 respostas válidas** e a maior parte das pessoas que responderam o survey são desenvolvedores (119 do total de respostas). Também foram recebidas respostas de gerentes de projetos, especialistas em segurança, arquitetos de software, entre outros. Em relação às questões demográficas, é possível notar que a maior parte dos participantes foi composta por homens, entre 21 e 30 anos de idade, sem crianças em casa, mas convivendo com alguma outra pessoa no mesmo ambiente. De forma geral, essas são as características demográficas dos participantes do survey 1:

Localização: Os participantes são de 38 diferentes países, com 20.7% do Brasil e 11.8% dos Estados Unidos.

Gênero: A maioria é formada por homens: 85.19%, enquanto 11.11% são mulheres e 3.70% preferiram não informar o seu gênero.

Idade: 42.96% dos participantes tem entre 21 e 30 anos de idade, 36.5% entre 31 e 40 anos, 10.2% tem menos de 21 anos, 8.90% entre 41 e 60 anos e 0.73% preferiram não informar a idade.

Estado civil: 57.7% dos participantes são solteiros, 36.29% são casados, e os outros são divorciados ou preferiram não fornecer essa informação.

Crianças: 75.56% dos participantes não vivem com crianças abaixo dos 12 anos de idade, 20.74% vivem com pelo menos uma criança abaixo de 12 anos e 3.70% não responderam essa questão.

Convívio: 78.52 % dos participantes vivem com pelo menos uma outra pessoa, 16.30% vivem sozinhos e 5.18% não responderam a essa questão.

Em relação à proximidade com o COVID-19, dois participantes (1,4%) informaram que testaram positivo para a doença e cinco (3,7%) informaram que alguém que vivia com eles foi diagnosticado com COVID-19. Além disso, 11 participantes (8.14%) relataram que um familiar ou amigo testaram positivo para a COVID-19.

A fim de investigar como a pandemia afetou a produtividade pessoal, os participantes foram solicitados a classificar a sua auto-produtividade antes e depois da pandemia. Inicialmente, foi solicitado que o participante classificasse sua produtividade pessoal antes das medidas de isolamento social em uma escala de 0 a 10, por meio da seguinte questão: “*How would you rate your usual contribution performance over the past year or two?*”. Em seguida os participantes foram indagados sobre sua produtividade durante o período de pandemia, por meio da seguinte questão “*How would you rate your overall contribution performance on the period you stayed at home due to the COVID-19 pandemic?*”—também usando uma escala de 0 a 10. Posteriormente, foi calculada a diferença entre as duas respostas para se verificar mudanças na percepção da própria produtividade. Essa abordagem para se estimar como os desenvolvedores avaliam os impactos da pandemia na sua produtividade foi baseada em um estudo anterior [6]. Os resultados dessa abordagem foram os seguintes: 26% dos participantes tiveram uma diminuição na percepção da produtividade, 42% não reportaram mudanças e 32% tiveram um aumento na produtividade reportada durante o período de medidas de restrição em função do período de isolamento social. Alguns comentários, ilustrados a seguir, refletem o contraste encontrado nesses números. Por exemplo, alguns participantes comentaram que o isolamento social teve um impacto positivo na sua produtividade:

“Mostly productivity is up for me because I’m not commuting any more. I can spend more time programming”.

“During the pandemic, I’m working from home, so I feel more comfortable spending the time to contribute upstream changes that I otherwise may not have during work hours”.

“I devoted myself more to working in OSS, because there were no distractions at home, such as constantly asking colleagues, and full access to the Internet appeared, since at work access to the Internet is very limited, which significantly slows down the work with github”.

De forma contrastante, alguns participantes reportaram dificuldades que implicaram diminuição da produtividade no período de isolamento social. Sendo, inclusive, algumas questões ligadas a questões familiares:

“I have less time for de open source contributions. I spend more time on my family”.

“The work related productivity is comparable, but it’s harder to find time for de open source projects. My family needs more attention now”.

Nesse sentido, foi verificado se os participantes que indicaram conviver com crianças com menos de 12 anos reportaram mudanças na produtividade: enquanto 27.4% das pessoas sem crianças reportaram um aumento na produtividade durante a pandemia, apenas 10.7% dos participantes com crianças nessa faixa de idade perceberam um aumento na produtividade. Além disso, a Figura 4.18 (a) mostra que a maioria dos participantes que vive com crianças abaixo de 12 anos reportou diminuição na produtividade durante a pandemia. Também é possível observar a partir da Figura 4.18 (b), que a maior parte das pessoas que reportaram não viver sozinha também teve uma diminuição na produtividade reportada.

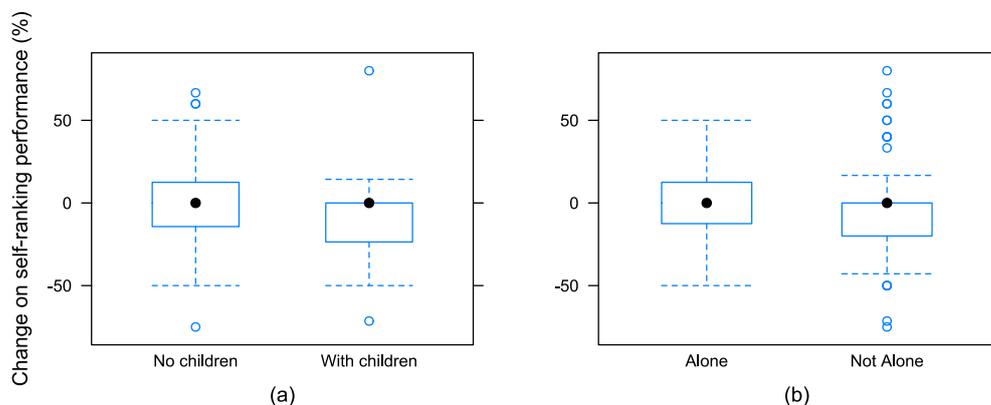


Figura 4.18: Mudanças na classificação da auto-produtividade dos participantes depois do início da pandemia, agrupados de acordo com o convívio familiar.

De maneira geral, a maioria (77.77%) dos participantes do survey 1 já faziam suas contribuições a partir de suas casa mesmo antes da pandemia. Contudo, as mudanças no ambiente com o advento da pandemia parecem também ter afetado esses contribuidores. A Figura 4.19 (a) demonstra a variação da percepção da própria produtividade entre os que já contribuía de casa antes da pandemia (Home) e os que faziam as contribuições de outro lugar (outside), como por exemplo a partir de um escritório. Embora exista uma maior concentração de pessoas que acreditam que houve aumento na autoprodutividade, a Figura 4.19 (b) indica que houve uma maior variação entre os participantes que já contribuía de casa mas que passaram a conviver com outras pessoas em isolamento durante a pandemia (Home and not alone), em relação aos contribuidores que já contribuía de casa, mas que passaram o período de isolamento sozinhos (Home and alone).

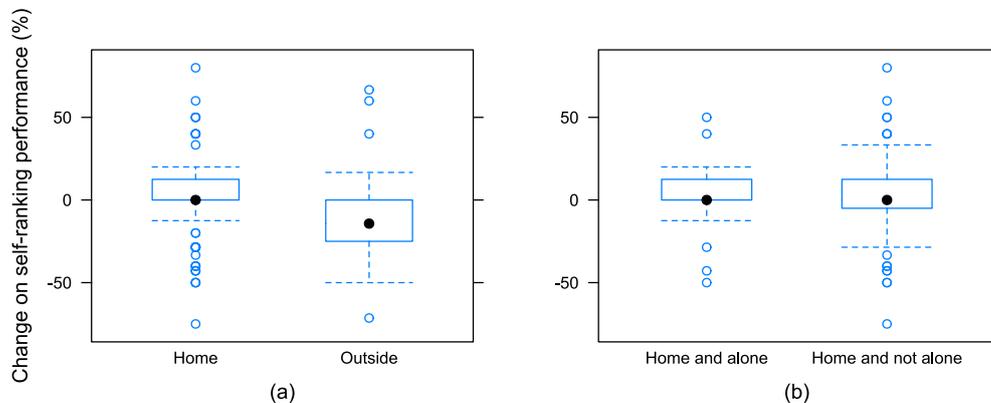


Figura 4.19: Mudanças na classificação da auto-produtividade dos participantes, agrupados de acordo com o local de onde os participantes contribuía antes da pandemia.

Os participantes também foram indagados “*How would you compare the contributions in the GitHub projects before and during the pandemics?*” a fim de se verificar suas percepções sobre a produtividade geral dos repositórios. A maioria deles responderam que as atividades dos repositórios mantiveram no mesmo nível (42.2%), 26.6% reponderam que havia mais contribuições durante a pandemia e 31.2% que havia mais contribuições antes da pandemia.

Os participantes também foram encorajados a indicar se a pandemia impactou seu estado emocional individual. Por meio das respostas recebidas é possível notar que a maioria dos contribuidores se sentiu cansado após as medidas de isolamento social (Figura 4.20), inclusive 42.2% dos participantes relataram exaustão. De forma geral, 70% dos participantes relataram cansaço ou exaustão. O sentimento de ansiedade também foi reportado por quase metade dos participantes. Algumas respostas ilustram a situação:

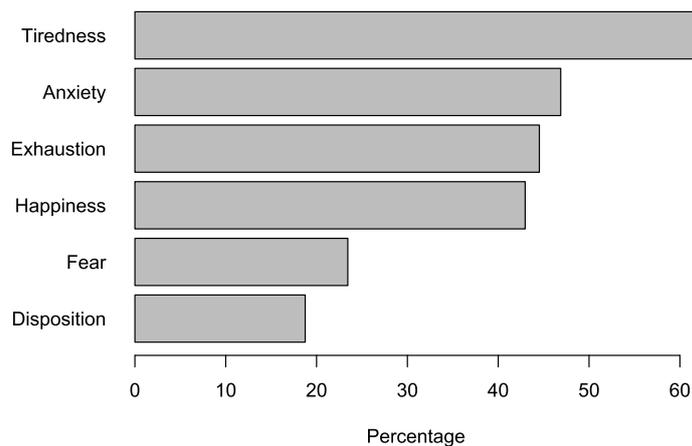


Figura 4.20: Sentimentos dos participantes durante as medidas de isolamento social.

"I feel anxious and exhausted all the time so everything is kind of overwhelming me".

"I never suffered from anxiety before, it was the first time I experienced this, but still at a low level".

Contudo, há algumas evidências que esses sentimentos não foram dominantes durante todo o período em que os contribuidores ficaram em casa por causa das medidas de isolamento, por exemplo, quando perguntados se eles concordavam com a afirmação *"During the period I stayed at home due to the COVID-19 pandemic, I have felt calm and relaxed"*, a maioria dos participantes reportaram que concordavam com a afirmação a maior parte do tempo e apenas 2 participantes (1.4%) indicaram que eles não concordavam com a afirmação em nenhum momento (Figura 4.21).

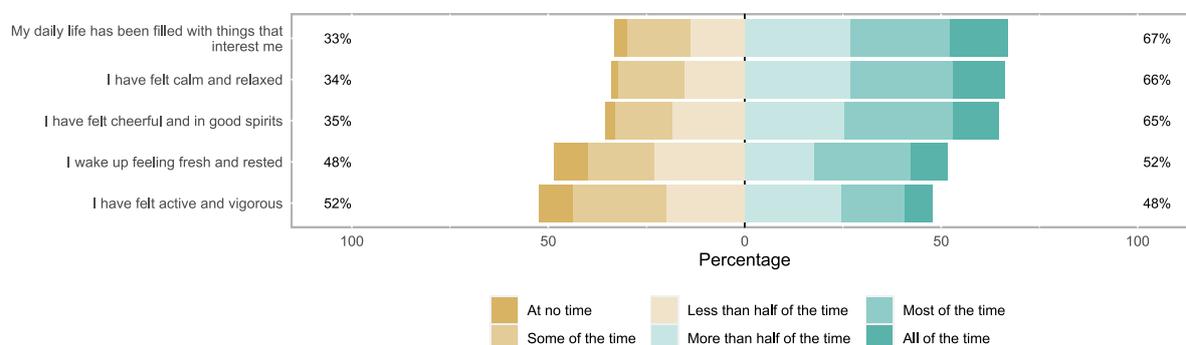


Figura 4.21: Como os participantes se sentiram durante o período de isolamento social.

Outras percepções foram exploradas usando o mesmo formato de questão (“*During the period I stayed at home due to the COVID-19 pandemic, I have felt ...*”), do mesmo modo que no estudo realizado por Ralph et al. [6]. A Figura 4.21 sintetiza as respostas recebidas. É possível notar que com exceção da afirmação “*During the period I stayed at home due to the COVID-19 pandemic, I have felt active and vigorous*” que na maior parte do tempo houve uma prevalência de sentimentos positivos durante o período que os participantes ficaram em isolamento social.

Observação 11. Os resultados sugerem que a pandemia da COVID-19 impactou os contribuidores de maneiras diferentes. Por exemplo, enquanto 32% acreditam que sua autoproductividade aumentou, 26% reportaram uma diminuição. Além disso, há evidências de que contribuidores vivendo sozinho ou sem crianças com idade abaixo de 12 anos reportaram uma maior produtividade. Por fim, foi encontrada uma alta frequência de sentimentos negativos durante a pandemia (70% reportaram cansaço/exaustão e 46.15% reportaram ansiedade).

Com o intuito de verificar se a pandemia teve efeitos sobre os sentimentos expressados nos comentários de pull requests e assim, identificar possíveis mudanças no estado emocional dos contribuidores, foi realizada a classificação da polaridade (negativo, neutro ou positivo) de todos os comentários de pull requests realizados nos repositórios do dataset, durante os 24 meses observados antes e durante a pandemia. No total foram analisados 2.661.886 comentários utilizando a ferramenta SentiSD [55] com o treinamento de um modelo utilizando uma base pública com mais de 7.000 documentos com anotações do GitHub [58]. A maior parte dos comentários foi classificada como neutra, o que é natural dado que esse tipo de discussão é predominante técnica [48]. Contudo, uma parte dos comentários foi classificado como tendo algum tipo de sentimento e, assim, a proporção em relação ao total de comentários foi calculada e agrupada em períodos mensais. A Figura 4.22 sintetiza os resultados e indica uma leve diminuição na proporção de comentários classificados como positivo durante os quatro primeiros meses da pandemia (meses de índice 1, 2, 3 e 4) e uma volta a patamares anteriores ao longo dos meses; mas não houve alterações perceptíveis na proporção de sentimentos negativos no período.

Observação 12. Foi percebida uma diminuição na proporção de comentários com polaridade classificada como positiva nos quatro primeiros meses da pandemia, com uma volta a patamares anteriores ao longo dos meses.

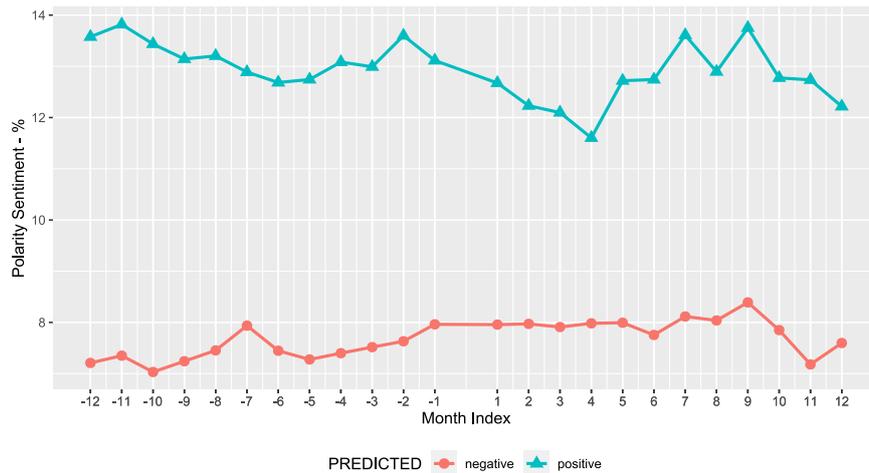


Figura 4.22: Polaridade dos sentimentos nos comentários de pull requests ao longo dos meses.

4.5 Síntese do Capítulo

Este capítulo apresentou resultados quantitativos, explorando os indicadores relacionados à atividade dos repositórios, ao ingresso de novos contribuidores e ao turnover dos core developers. Foram exploradas as diferenças entre o nível de atividade e as tendências desses indicadores após o início da pandemia. Também foram apresentados os resultados qualitativos, considerando a produtividade reportada pelo participantes e questões relacionadas às suas percepções decorrentes da pandemia.

Capítulo 5

Análise dos Resultados

Este capítulo apresenta as análises dos estudos quantitativos e qualitativos realizados para se entender os impactos da pandemia no nível de atividade dos repositórios e no bem-estar e estado emocional dos contribuidores durante a pandemia da COVID-19. Em seguida, são comparados os resultados dessa dissertação com estudos correlatos.

5.1 Análises do esforço de mineração

De forma geral os resultados do esforço de mineração indicam que houve um aumento na mediana de commits, pull requests e comentários após o início da pandemia. Curiosamente, a mediana do commit churn permaneceu a mesma nos dois períodos, o que indica que não houve diminuição no esforço de cada contribuição individual. Dessa forma, há indícios de que os desenvolvedores produziram mais, mas sem diminuir o tamanho das contribuições. A percepção dos core developers vai ao encontro dessa constatação, já que a maioria dos participantes do survey 2 que percebeu alguma variação respondeu que houve aumento no nível de todos os indicadores, após o início da pandemia. Contudo, os resultados do esforço de mineração sugerem que houve um aumento expressivo de contribuições apenas no início da pandemia - os gráficos e os modelos de regressão RDD indicam que esse aumento foi determinante para que a mediana dos indicadores no período durante a pandemia fosse maior do que a mediana do período anterior. Por exemplo, os 155 repositórios analisados tiveram uma mediana de 147 pull requests no primeiro mês, esse valor é quase 25% maior do que a média de todo o período de 24 meses observado. Portanto, a percepção de um aumento do nível de atividades por parte dos desenvolvedores e o aumento da mediana percebido a partir dos indicadores podem estar relacionados ao aumento ocorrido no início da pandemia. Além disso, tanto os gráficos, quanto os coeficientes dos modelos RDD sugerem uma volta a patamares anteriores ao longo dos meses.

O aumento de atividade percebido de maneira geral, pode não refletir a situação completa. Há de se notar que a maior parte dos repositórios não apresentou uma diferença estatística significativa comparando o antes e o durante. E o mesmo ocorre com a percepção dos core developers durante o survey 2: a maioria reportou que de modo geral não houve diferença no nível de atividades por causa da pandemia. Além disso, os resultados apresentam uma contradição, pois considerando os repositórios que apresentaram diferença significativa entre os dois períodos, o efeito foi negativo na maioria deles; por exemplo, dos 76 repositórios que tiveram diferença significativa no número de pull requests, 50 tiveram um efeito negativo e apenas 26 tiveram um efeito positivo. E essa constatação não está totalmente alinhada com a percepção dos core developers, já que uma menor parte deles reportou diminuição em todos os indicadores, incluindo o número de pull requests. A Tabela 5.1 sintetiza alguns resultados: em relação ao MSR, são considerados neutros (N) os repositórios que não tiveram diferença estatística significativa comparando os dois períodos analisados e observada a direção do tamanho do efeito (positivo ou negativo) para considerar aumento ou diminuição no nível de atividade; já em relação ao Survey 2, a Tabela 5.1 resume a percepção dos usuários em relação a cada um dos indicadores analisados. A combinação dos dados na tabela evidencia a contradição mencionada, com um maior número de repositórios com diminuição nos indicadores, enquanto a maior parte dos core developers que relatou alterações indicou que a alteração foi positiva.

| | MSR | | | Survey 2 | | |
|--------------|-----|-----|-----|----------|-----|-----|
| | - | N | + | - | N | + |
| Latência PRs | 14% | 71% | 15% | 20% | 44% | 36% |
| PRs | 32% | 51% | 17% | 16% | 50% | 34% |
| Commits | 28% | 51% | 21% | 18% | 48% | 34% |
| Comentários | 23% | 61% | 16% | 16% | 56% | 28% |
| Churn | 16% | 76% | 8% | 22% | 56% | 22% |

Tabela 5.1: O resumo dos resultados do esforço de MSR e do Survey 2 em relação ao nível de atividade dos repositórios

Dessa forma, com base nos resultados não se pode afirmar que a pandemia impactou as atividades de desenvolvimento nos projetos OSS de forma uniforme (com predominância de aumento ou diminuição no nível de atividade). Essa mesma constatação foi observada no estudo conduzido por Silveira et al. [14], os autores consideram que o impacto da pandemia em projetos OSS não foi binário (aumento ou diminuição da produtividade), mas um spectrum. A justificativa dos autores é que muitas das observações realizadas resultam em opiniões divergentes sobre o efeito da pandemia. Essa constatação está de alguma forma alinhada com os resultados dessa dissertação; por exemplo, 28% dos repo-

sitórios apresentaram uma diminuição no número de commits e 21% apresentaram um aumento. Esses resultados indicam que os repositórios foram impactados de maneiras diferentes, mas o balanço final foi positivo em função do aumento dos indicadores observado do início da pandemia.

O turnover de core developers também teve o maior valor captado durante o início da pandemia. Embora intuitivamente pareça ser um problema, o aumento do turnover nesse período pode não ter causado danos expressivos. Esse aumento pontual pode estar relacionado a uma maior quantidade de contribuições realizadas por pessoas que se tornaram ou voltaram a ser core developers nesse período. Essa constatação é sustentada por uma maior taxa de Core Developers Newcomers em relação aos Core Developers Leavers e também pelo baixo número de desengajamento dos core developers: de um total de 582 core developers existentes no primeiro mês da pandemia, apenas 50 (8.5%) não voltaram a esse posto ao longo dos meses (mas continuaram contribuindo) e apenas 8 (1.3%) deles abandonaram o projeto (não fizeram nenhuma contribuição ao longo dos meses). Esse desengajamento, pode ser um efeito natural, já que estudos anteriores [39] indicam que 33% dos core developers costumam desengajar dos projetos pelo menos uma vez. Além disso, tanto a média de turnover mensal quanto o turnover anual apresentaram diminuição quando comparados a períodos anteriores.

Já o outro extremo, a entrada de novos contribuidores, apresentou resultados mais preocupantes. Tanto a análise mês a mês, quanto a análise anual, sugerem uma diminuição no ingresso de novos contribuidores. E essa diminuição parece ter atingido grande parte dos repositórios observados, já que, durante a pandemia, 63% deles tiveram uma redução no ingresso de novos contribuidores, sendo os projetos escritos nas linguagens de programação Javascript e Go os que mais sofreram perdas de novos contribuidores.

Em relação ao engajamento dos novos contribuidores, não foram percebidas variações expressivas. Os dados sugerem que a proporção de pessoas que fizeram apenas uma contribuição e não voltaram a contribuir durante o ano manteve-se estável. Contudo, essa constatação não atenua os efeitos causados pela diminuição no ingresso de novos contribuidores e pode levar a impactos no médio e longo prazo.

5.2 Análises dos surveys

O survey 1 foi direcionado a contribuidores de projetos de OSS de modo geral, e teve como intuito capturar as percepções dos desenvolvedores sobre alterações no nível de atividade dos repositórios, capturar a percepção sobre a autoprodutividade e também sobre o bem-estar e estado emocional durante o período de pandemia. Já o survey 2 foi destinado apenas aos desenvolvedores considerados core developers e teve como objetivo identificar

a visão desse grupo restrito de contribuidores sobre o impacto da pandemia nas atividades dos repositórios.

Em relação à percepção dos participantes dos dois surveys sobre o nível de contribuições gerais dos repositórios, antes e depois da pandemia, os resultados são consistentes: a maior parte considerou que não houve aumento e houve uma divisão entre os que perceberam algum aumento, sendo que uma maior proporção considerou que antes da pandemia havia mais contribuições, conforme dados da Tabela 5.2.

| | Survey 1 | Survey 2 |
|--|----------|----------|
| Before the pandemics there was more contributions. | 31.2% | 31.6% |
| The level of contributions remain the same. | 42.2% | 47.4% |
| During the pandemics there was more contributions. | 26.6% | 21.1% |

Tabela 5.2: Os resultados do Survey 1 e do Survey 2 em relação a percepção dos participantes sobre o nível de atividade dos repositórios.

Já em relação à percepção da autoproductividade, no survey 1 a maior parte dos participantes (43%) reportou que a pandemia não acarretou mudanças significativas na sua própria produtividade. Por outro lado, 32% reportou aumento na produtividade e 25% reportou diminuição. No survey 2, 38.6% dos participantes indicou estabilidade, 31.6% reportou um aumento e 29.8% indicou que a produtividade diminuiu durante a pandemia. Dessa forma, novamente a percepção da produtividade está alinhada entre os participantes dos dois surveys, com uma maior parte indicando que a produtividade manteve-se a mesma e com uma divisão entre os que acreditaram que houve mudanças, sendo uma maior parte formada pelos que acreditam em um aumento da própria produtividade. Essa percepção da própria produtividade está de acordo com o segundo survey realizado por Ford et al. [59], no qual os autores também reportam que mais profissionais (36%) da Microsoft, que fizeram parte desse survey, reportou um aumento na sua produtividade após o início da pandemia, contrastando com os profissionais que reportaram diminuição (30%).

Contudo, há indícios de que alguns grupos foram mais afetados pela pandemia. Os resultados do survey 1 sugerem que pessoas que moram sozinhas e/ou que não tenha em casa crianças com idade inferior a 12 anos tiveram uma percepção de produtividade maior do que pessoas que convivem com familiares, companheiros de quarto e crianças. De fato, o estudo de Ford et al. [59] indica a proximidade com membros da família em função das medidas de isolamento social como um dos desafios mais frequentes reportados - muito por causa dessa proximidade ser uma fonte de interrupções e distrações.

Sobre o estado emocional e bem-estar dos contribuidores, foram percebidos maiores relatos sobre sentimentos negativos, com uma alta frequência de sentimentos negativos

durante a pandemia (70% reportou cansaço/exaustão e 46.15% reportou ansiedade). Contudo, alguns fatores indicam que esses sentimentos não preencheram todo o tempo dos participantes, pois o terceiro sentimento mais citado foi justamente a felicidade. Além disso, mais da metade dos participantes indicou que se sentiu ativo e vigoroso pelo menos mais da metade do tempo.

5.3 Discussões e Limitações

Os resultados dessa dissertação indicam um aumento substancial nas contribuições (commits, pull requests e comentários) no início da pandemia, mas que esse aumento não se tornou uma tendência ao longo do tempo. Essa constatação está alinhada com o GitHub Octoverse 2020 Productivity Report [12], cujo resultado sugere que o tempo entre o primeiro e o último pull request e a quantidade de trabalho feito durante o dia aumentou nos primeiros meses da pandemia em diversas regiões e então estabilizou ao longo dos meses. Contudo, o relatório do GitHub investigou o número de pull requests por desenvolvedor, enquanto nessa dissertação são exploradas as atividades de desenvolvimento no nível dos repositórios.

Outro estudo que teve como foco o desenvolvimento de OSS durante a pandemia foi realizado por Silveira et al. [14], nesse caso os resultados também estão alinhados de alguma forma. Embora os períodos observados e as metodologias empregadas sejam diferentes, os autores encontraram um aumento de pull requests abertos, comparando os períodos de Julho de 2019 a Dezembro de 2019 e Janeiro de 2020 a Maio de 2020. Esse segundo período engloba o primeiro mês da pandemia considerado nessa dissertação, em que realmente houve o aumento mais expressivo nos indicadores observados, incluindo os pull requests. Além desse estudo, Wang et al. [28] também exploraram o nível de atividade nos repositórios durante a pandemia, e novamente, há um alinhamento entre os resultados: os autores perceberam um aumento na soma de contribuições durante o início da pandemia (compreendendo o período do meio de março até o final de junho). Esse período de crescimento também engloba o período do pico de contribuições detectados nessa dissertação.

Dessa forma, embora os repositórios e os participantes dos surveys dessa dissertação tenham sido escolhidos a partir de uma amostra, o alinhamento com os relatórios do GitHub (que analisa toda a população de desenvolvedores) e com os outros estudos que estudaram os efeitos da pandemia no ecossistema OSS, indicam que alguns resultados, como o pico de contribuições no início da pandemia, possam também se aplicar de forma geral. A Tabela 5.3 resume os quatro estudos (incluindo essa dissertação). Os principais

resultados foram sintetizados para indicar os alinhamentos e contradições com a presente dissertação.

| | GitHub [12] | Silveira et al. [14] | Wang et al. [28] | Essa dissertação |
|-----------------------|---|--|--|--|
| Objetivo | Investigar como o ritmo de trabalho e atividade mudaram durante a pandemia. | Investigar o impacto da pandemia em projetos e profissionais de software. | Investigar o impacto da COVID-19 na atividade e contribuições de desenvolvedores de OSS. | Investigar impactos no processo de desenvolvimento e no bem-estar de desenvolvedores de OSS |
| Método | Análise descritiva das atividades de desenvolvedores do GitHub | MSR e survey | MSR | MSR e dois surveys |
| População | Todos os projetos de código aberto e pago | 100 repositórios em Java do GitHub e 279 profissionais | Todos os usuários do GitHub | 155 repositórios de diferentes linguagens, 137 desenvolvedores e 57 core developers |
| Principais resultados | Atividades consistentes com anos anteriores; mudanças no ritmo de trabalho; aumento na janela entre o primeiro e o último push no início da pandemia. | Ingresso de novos contribuidores diminuíram; WFH não impactou as contribuições, produtividade e tempo para completar tarefas; distúrbios do sono e stress aumentaram e felicidade diminuiu; interrupções não impactaram negativamente. | Aumento significativo nas contribuições no início da pandemia; nível de atividade correlacionadas com o surto de Covid; repositórios afetados de modo diferente. | Aumento de atividade e turnover de core developers no início da pandemia; diminuição no ingresso de novos contribuidores; predominância de sentimentos negativos; pessoas sem crianças e sozinhas tem melhor percepção da produtividade. |

Tabela 5.3: Os resultados dos estudos relacionados ao desenvolvimento de software de código aberto.

Considerando os 155 repositórios, o turnover dos core developers teve grandes alterações no início da pandemia. Mas de maneira geral, esse período não apresentou grandes impactos no abandono dos core developers. Tanto a média de turnover mensal, quanto o turnover anual tiveram reduções quando comparados com períodos anteriores. Esses resultados estão de alguma forma alinhados com [1], cujo estudo foi realizado em tempos normais e cujos resultados sugerem que a maioria dos repositórios tem um turnover anual de 30% a 40%. Realmente, nessa dissertação a média anual de turnover detectada nos repositórios observados durante a pandemia foi de 32.17%. Contudo, nessa dissertação foi detectada uma alteração na distribuição do turnover entre os repositórios quando observados os níveis de atividade: antes da pandemia repositórios com menos commits possuíam maior turnover e após repositórios com mais commits apresentaram um maior turnover. Embora não tenha sido detectada uma correlação entre nível de atividade e turnover, estudos futuros devem ser conduzidos para entender os efeitos a longo prazo.

Outro ponto que pode provocar efeitos a longo prazo é a diminuição no ingresso de novos contribuidores percebidos ao longo de toda a pandemia. A maior parte dos repositórios sofreu com a diminuição no ingresso e o ingresso total foi menor no ano da pandemia do que nos três anos anteriores. Esses resultados estão de alguma forma alinhados com o estudo de Silveira et al. [14], que também explorou o ingresso de novos contribuidores durante a pandemia e encontrou diferença significativa nesse indicador comparando os períodos de Janeiro de 2020 a Março de 2020 e Abril de 2020 a Maio de 2020. No primeiro período o ingresso era de 8.88 contribuidores em média e no segundo, a média caiu para 4.46. Além disso, os autores relatam que 70% dos repositórios diminuíram o ingresso de novos contribuidores no mesmo período. Contudo, os autores analisaram apenas projetos escritos em Java, enquanto nessa dissertação foram exploradas outras linguagens. Sendo inclusive percebida diferenças entre elas: entre os projetos que tiveram diminuição no ingresso, as maiores foram percebidas por repositórios escritos nas linguagens Javascript e Go. Embora seja difícil entender o comportamento de pessoas que sequer chegaram a contribuir, novos estudos podem ser conduzidos focando, por exemplo, em usuários que realizaram um fork dos repositórios e que não chegaram a realizar um pull request no repositório original.

Em relação à percepção dos desenvolvedores sobre a própria produtividade, a maior parte deles indica que não houve mudanças, e mesmo quando acreditam que a pandemia afetou seu desempenho, existe uma dualidade nas respostas. Dessa forma, é possível que a pandemia que ocorreu de forma inesperada e abrupta tenha tido efeitos diferentes nas pessoas. E esses efeitos podem ter causas diversas, por exemplo os participantes dos surveys citaram diversos fatores que podem ter causado efeito positivo na produtividade: não-necessidade de locomoção, menos interrupções de colegas de trabalho, internet

mais rápida e sem bloqueios. Outros citaram fatores negativos: interrupções por outros membros da família, problemas de saúde decorrentes da COVID-19 e fatores psicológicos. Ford et al. [59] chamou esse efeito de Yin e Yang, em função da dicotomia encontrada nos resultados durante a pandemia: muitas vezes um mesmo fator pode ter efeito positivo e negativo para pessoas diferentes, como por exemplo a proximidade com a família.

Nas questões relacionadas a como os participantes se sentiram durante a pandemia, os resultados do survey 1 sugerem uma prevalência de sentimentos negativos (70% reportou cansaço ou exaustão e 46.15% reportou ansiedade). Os resultados de [8], que investigou os fatores que influenciaram a produtividade no desenvolvimento de software no Brasil durante a pandemia, também encontrou uma prevalência de sentimentos negativos: o sentimento mais comum também sendo ansiedade e cansaço/exaustão. Contudo, os resultados do survey 1 indicam a felicidade como o terceiro sentimento mais reportado, enquanto em [8] esse foi o menos citado. Ainda em relação aos sentimentos, foi percebida uma diminuição na proporção de comentários de pull requests com polaridade classificada como positiva durante os primeiros meses da pandemia; esse período coincide com o aumento na frequência dos comentários, conforme reportado nas análises do esforço de mineração do Capítulo 4. Dessa forma, um trabalho futuro pode ser conduzido para verificar se a diminuição na proporção de comentários positivos ocorreu em função da alteração na quantidade de comentários ou se foi um efeito direto da pandemia.

Os resultados de Ralph et al. [6] indicam que a produtividade e o bem-estar foram prejudicados durante a pandemia. Quanto à produtividade, os resultados dessa dissertação diferem do resultado deles, pois os autores relatam que mais contribuidores reportaram aumento do que diminuição na percepção da produtividade durante a pandemia. De forma contrastante tanto no survey 1, quanto no survey 2 mais participantes reportaram uma diminuição do que aumento da autoprodutividade. Apesar disso, os resultados dessa dissertação indicam uma alta frequência de sentimentos negativos entre os contribuidores, o que está de alguma forma alinhado os resultados de [6].

Por fim, tanto o nível de atividade dos repositórios, quanto a percepção dos contribuidores sobre o feito da pandemia tiveram uma dualidade, com consequências que devem ser observadas a longo prazo. Tanto os estudos que exploraram o impacto da pandemia nos projetos OSS [14], quanto os estudos realizados em empresas privadas [59], concordam que não existe apenas uma face da moeda. Os efeitos parecem ter sido sentidos de formas diferentes e isso causou reações diversas nas pessoas, incluindo o comportamento em relação ao desenvolvimento de software livre. Alguns aspectos exigem mais cuidado, especialmente se a diminuição no fluxo de novos desenvolvedores irá prejudicar de alguma forma a continuidade dos projetos. Mas a pandemia parece ter criado oportunidades para explorar melhorias nos projetos e desafios que devem ser superados. Outro trecho menciona

onado por Ford et al. [59], retirado do livro *Um Conto de Duas Cidades* resume bem o período “It was the best of times, it was the worst of times.”

5.4 Ameaças para Validação

Nesse trabalho foi analisado parte do histórico de desenvolvimento de 155 repositórios, vindos do dataset inicial contendo 1500 repositórios. Esses 155 repositórios foram selecionados usando um critério de inclusão baseado na quantidade de commits, pull requests e número de contribuidores. A razão para se filtrar os repositórios dessa maneira é utilizar apenas projetos populares, ativos e que realmente sejam um repositório de software. Embora a amostra selecionada seja diversa, em termos de atividade, linguagem de programação e número de contribuidores, os resultados não podem ser generalizados a sistemas que não satisfaçam o critério de inclusão. Ainda em relação à validade externa, o survey 1 com os desenvolvedores foi respondido por 135 participantes e o survey 2 por 57 core developers, de forma que essa quantidade poderia não ser representativa o bastante para fazer conclusões definitivas. Contudo, os resultados do esforço de mineração e do survey se sustentam e dessa forma, os resultados do survey podem ser utilizados para entender como os desenvolvedores acreditam que a pandemia influenciou sua produtividade e engajamento com o desenvolvimento de OSS.

Com o intuito de convidar desenvolvedores a responder o questionário, inicialmente foram extraídos os endereços de e-mail dos contribuidores diretamente dos artefatos do GitHub. Posteriormente, esses endereços foram usados para enviar mensagens de convite aos contribuidores para participação do survey. Embora exista algumas preocupações de privacidade em relação a essa prática, esse método foi escolhido para se abranger uma amostra maior. Vale a pena mencionar que a GitHub Privacy Statements ¹ permite que pesquisadores acessem os dados dos usuários e que, após o envio dos e-mails, todos os endereços sejam excluídos para inibir que esses dados sejam usados no futuro para outro fim. Essa abordagem não impede a reprodução desse estudo, pois os e-mails foram usados apenas para o processo de convite dos participantes. Além disso, foram recebidos três respostas positivas de participantes em relação aos e-mails enviados. Um deles disse que estava interessado no resultado da pesquisa e solicitou que o trabalho fosse enviado, quando terminado. Outro participante inclusive se ofereceu para divulgar o questionário. Dessa forma, o método de envio adotado nesse estudo não parece ter causado nenhum dano à privacidade dos desenvolvedores.

Inicialmente foi utilizada uma heurística baseada na contagem de commits para computação dos core developers. Assim, foi tomado como premissa que os core developers

¹<https://docs.github.com/en/github/site-policy/github-privacy-statement>

são os autores que contribuíram com 80% dos commits no período de um ano. Essa abordagem pode não capturar outras atividades importantes como a mentoria, reuniões e acolhimento de novos desenvolvedores. Para tentar diminuir essa ameaça também foi realizada as análises utilizando outro algoritmo de identificação dos principais desenvolvedores: o TruckFactor [47]. Contudo, mesmo utilizando as duas abordagens é possível que possam faltar outros contribuidores que fizeram contribuições significativas aos projetos. Esse problema tem origem na falta de consenso sobre quais fatos caracterizam os core developers em um projeto, e assim foi adotada definições específica da literatura [1] e [47]. De forma similar para estimar o nível de atividade dos repositórios alguns indicadores foram utilizados nesse estudo; contudo, outros indicadores também poderiam ser adotados, mas nessa dissertação foram utilizados métricas já usadas como representantes do nível de atividades em estudos anteriores [20, 32, 33].

Em relação à classificação da polaridade dos sentimentos dos comentários de pull requests, não foi aferida a performance do modelo de classificação. Contudo, a abordagem utilizada foi baseada em um estudo anterior [58], com o treinamento do modelo com a mesma base de dados e que obteve índices satisfatórios. O recall, precision e F1 para as classes de polaridade são respectivamente: negativo (.92, .90, .91), neutro (.90 .93 .92) e positivo (.95 .91 .93). Dessa forma, espera-se que os resultados do modelo de classificação utilizado nessa dissertação tenha desempenho similar ao estudo de [58].

Capítulo 6

Conclusões

Os resultados dessa dissertação indicam que a maioria dos contribuidores de projetos OSS já executavam suas atividades de casa mesmo antes da pandemia, mas ainda assim, as contribuições foram impactadas, especialmente nos primeiros dias das medidas de isolamento. Foi detectado um aumento significativo de pull requests, commits, comentários e de turnover de core developers logo no início da pandemia. Contudo, os resultados sugerem que esses indicadores estabilizaram ao longo dos meses. Além disso, há indícios de que a pandemia da COVID-19 não afetou os contribuidores da mesma forma, com parte deles reportando um aumento e parte reportando uma diminuição no nível de produtividade. Essa dualidade apresentou efeitos também no nível de atividade dos repositórios, onde também foi percebida uma divisão entre os projetos que apresentaram aumento ou diminuição nos indicadores. Essas diferenças dos efeitos da pandemia podem estar ligadas a algumas características, pois durante o survey 1 foi percebido um maior aumento da produtividade em pessoas que moravam sozinhas ou que não conviviam com crianças; além disso, projetos escritos em Javascript e Go tiveram uma maior diminuição no ingresso de novos contribuidores. Outros fatores, porém, foram mais abrangentes, dado que a maioria dos contribuidores reportou cansaço ou exaustão durante o período de isolamento social. A diminuição no ingresso de novos contribuidores também é um fator que necessita de atenção, tanto a série histórica, quanto a análise mês a mês mostram uma diminuição na média de ingresso; além disso, a maioria dos repositórios teve diminuição na média de ingresso durante a pandemia. Dessa forma, é possível verificar que a pandemia trouxe oportunidades para aumentar a produtividade dos contribuidores e, conseqüentemente, o nível de atividade dos repositórios, mas também a necessidade de se tomar medidas a ajudar os desenvolvedores nesse período de incerteza.

6.1 Trabalhos Futuros

Como trabalho futuro pretende-se investigar quais as implicações a longo prazo dos efeitos causados pela pandemia nos projetos OSS - não está claro como o alto grau de turnover observado no início da pandemia e como a diminuição de novos contribuidores nos projetos afetou o nível de qualidade ou mesmo a continuidade dos sistemas. Outro ponto a ser estudado são os impactos na qualidade dos projetos, inclusive os modelos de regressão desse estudo podem ser ampliados adicionando indicadores de qualidade para se entender se houve impacto na qualidade dos projetos e quais fatores tiveram maior efeito nesse tipo de indicador, após o início da pandemia.

Também pretende-se investigar como a pandemia afetou grupos específicos em perspectivas relacionadas à diversidade dos times, como por exemplo, o caso das mulheres que contribuem com projetos OSS. As mulheres podem ter sido impactadas de maneira diferente. Inclusive, alguns estudos já foram realizados para identificar impactos da pandemia na produtividade sob o ponto de vista do gênero, mas esses estudos tiveram o foco em áreas diferentes, sendo importante um estudo específico na área de projetos OSS.

Referências

- [1] Ferreira, Fabio, Luciana Lourdes Silva e Marco Tulio Valente: *Turnover in open-source projects: The case of core developers*. Em *Proceedings of the 34th Brazilian Symposium on Software Engineering*, páginas 447–456, 2020. ix, 13, 14, 26, 27, 32, 48, 49, 51, 52, 66, 69
- [2] Singhal, Tanu: *A review of coronavirus disease-2019 (covid-19)*. *The indian journal of pediatrics*, 87(4):281–286, 2020. 1
- [3] Kraus, Sascha, Thomas Clauss, Matthias Breier, Johanna Gast, Alessandro Zardini e Victor Tiberius: *The economics of covid-19: initial empirical evidence on how family firms in five european countries cope with the corona crisis*. *International Journal of Entrepreneurial Behavior & Research*, 2020. 1
- [4] Bao, Lingfeng, Tao Li, Xin Xia, Kaiyu Zhu, Hui Li e Xiaohu Yang: *How does working from home affect developer productivity? – a case study of baidu during covid-19 pandemic*, 2020. 1, 2, 10, 16
- [5] Mendonça, Walter Lucas Monteiro de, Pedro Henrique Teixeira Costa, Emille Catarine Rodrigues Cançado, Fernanda Lima, Edna Dias Canedo, Rodrigo Bonifácio e Luis Henrique Vieira Amaral: *From dusk till dawn: Reflections on the impact of covid-19 on the development practices of a r&d project*. Em *Proceedings of the 34th Brazilian Symposium on Software Engineering*, páginas 596–605, 2020. 1, 10, 18
- [6] Ralph, Paul, Sebastian Baltes, Gianisa Adisaputri, Richard Torkar, Vladimir Kovalenko, Marcos Kalinowski, Nicole Novielli, Shin Yoo, Xavier Devroey, Xin Tan, Minghui Zhou, Burak Turhan, Rashina Hoda, Hideaki Hata, Gregorio Robles, Amin Milani Fard e Rana Alkadhi: *Pandemic programming: How covid-19 affects software developers and how their organizations can help*, 2020. 1, 2, 7, 12, 16, 32, 54, 58, 67
- [7] Oliveira, José Nilmar Alves de, Jaime Orrillo e Franklin Gamboa: *The home office in times of covid-19 pandemic and its impact in the labor supply*, 2020. 1
- [8] Bezerra, Carla IM, José Cezar de Souza Filho, Emanuel F Coutinho, Alice Gama, Ana Livia Ferreira, Gabriel Leitão de Andrade e Carlos Eduardo Feitosa: *How human and organizational factors influence software teams productivity in covid-19 pandemic: A brazilian survey*. Em *Proceedings of the 34th Brazilian Symposium on Software Engineering*, páginas 606–615, 2020. 1, 10, 17, 67

- [9] Vasilescu, Bogdan, Alexander Serebrenik e Vladimir Filkov: *A data set for social diversity studies of github teams*. Em *2015 IEEE/ACM 12th working conference on mining software repositories*, páginas 514–517. IEEE, 2015. 1
- [10] Hertel, Guido, Sven Niedner e Stefanie Herrmann: *Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel*. *Research policy*, 32(7):1159–1177, 2003. 1
- [11] Russo, Daniel, Paul PH Hanel, Seraphina Altnickel e Niels van Berkel: *The daily life of software engineers during the covid-19 pandemic*. arXiv preprint arXiv:2101.04363, 2021. 2, 10, 17
- [12] GitHub: *Github octoverse 2020 productivity report*, 2020. <https://octoverse.github.com/>, (Date last accessed 01-April-2021). 2, 3, 10, 11, 19, 64, 65
- [13] GitHub: *Octoverse spotlight: An analysis of developer productivity, work cadence, and collaboration in the early days of covid-19*, 2019. <https://github.blog/2020-05-06-octoverse-spotlight-an-analysis-of-developer-productivity-work-cadence> (Date last accessed 01-April-2021). 2, 10, 11, 19
- [14] Neto, Paulo Anselmo da Mota Silveira, Umme Ayda Mannan, Eduardo Santana de Almeida, Nachiappan Nagappan, David Lo, Pavneet Singh Kochhar, Cuiyun Gao e Iftekhar Ahmed: *A deep dive on the impact of covid-19 in software development*. arXiv preprint arXiv:2008.07048, 2020. 2, 10, 11, 20, 61, 64, 65, 66, 67
- [15] Canedo, Edna Dias, Rodrigo Bonifácio, Márcio Vinicius Okimoto, Alexander Serebrenik, Gustavo Pinto e Eduardo Monteiro: *Work practices and perceptions from women core developers in oss communities*. Em *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, páginas 1–11, 2020. 2, 13, 24
- [16] Mockus, Audris: *Organizational volatility and its effects on software defects*. Em *Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering*, páginas 117–126, 2010. 2, 3, 43
- [17] Steinmacher, Igor, Igor Wiese, Ana Paula Chaves e Marco Aurélio Gerosa: *Why do newcomers abandon open source software projects?* Em *2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, páginas 25–32. IEEE, 2013. 2, 3, 12, 44
- [18] Kalliamvakou, Eirini, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M German e Daniela Damian: *The promises and perils of mining github*. Em *Proceedings of the 11th working conference on mining software repositories*, páginas 92–101, 2014. 3, 6, 7, 42
- [19] Peterson, Kevin: *The github open source development process*. url: <http://kevinp.me/github-process-research/github-processresearch.pdf> (visited on 05/11/2017), 2013. 3

- [20] Vasilescu, Bogdan, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu e Vladimir Filkov: *Gender and tenure diversity in github teams*. Em *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, páginas 3789–3798, 2015. 3, 11, 12, 69
- [21] Ye, Yunwen e Kouichi Kishida: *Toward an understanding of the motivation of open source software developers*. Em *25th International Conference on Software Engineering, 2003. Proceedings.*, páginas 419–429. IEEE, 2003. 3
- [22] Hemmati, Hadi, Sarah Nadi, Olga Baysal, Oleksii Kononenko, Wei Wang, Reid Holmes e Michael W Godfrey: *The msr cookbook: Mining a decade of research*. Em *2013 10th Working Conference on Mining Software Repositories (MSR)*, páginas 343–352. IEEE, 2013. 6, 7, 21, 22, 23, 28
- [23] Poncin, Wouter, Alexander Serebrenik e Mark Van Den Brand: *Process mining software repositories*. Em *2011 15th European Conference on Software Maintenance and Reengineering*, páginas 5–14. IEEE, 2011. 6
- [24] Hassan, Ahmed E: *The road ahead for mining software repositories*. Em *2008 Frontiers of Software Maintenance*, páginas 48–57. IEEE, 2008. 7
- [25] AlMarzouq, Mohammad, Abdullatif AlZaidan e Jehad AlDallal: *Mining github for research and education: challenges and opportunities*. *International Journal of Web Information Systems*, 2020. 7, 21, 22
- [26] Miller, Courtney, Paige Rodeghero, Margaret Anne Storey, Denae Ford e Thomas Zimmermann: *"how was your weekend?" software development teams working from home during covid-19*. arXiv preprint arXiv:2101.05877, 2021. 10, 18
- [27] Neumann, Michael, Yevgen Bogdanov, Martin Lier e Lars Baumann: *The sars-cov-2 pandemic and agile methodologies in software development: A multiple case study in germany*. Em *International Conference on Lean and Agile Software Development*, páginas 40–58. Springer, 2021. 10, 19
- [28] Wang, Liu, Ruiqing Li, Jiaxin Zhu, Guangdong Bai, Weihang Su e Haoyu Wang: *Understanding the impact of covid-19 on github developers: A preliminary study*. Em *33rd International Conference on Software Engineering and Knowledge Engineering, SEKE 2021*, páginas 249–254, 2021. 10, 11, 19, 64, 65
- [29] Oliveira, Edson, Eduardo Fernandes, Igor Steinmacher, Marco Cristo, Tayana Conte e Alessandro Garcia: *Code and commit metrics of developer productivity: a study on team leaders perceptions*. *Empirical Software Engineering*, 25(4):2519–2549, 2020. 10, 11
- [30] Tangen, Stefan: *Understanding the concept of productivity*. Em *Proceedings of the 7th Asia-Pacific Industrial Engineering and Management Systems Conference, Taipei*, páginas 18–20, 2002. 10

- [31] Vasilescu, Bogdan, Yue Yu, Huaimin Wang, Premkumar Devanbu e Vladimir Filkov: *Quality and productivity outcomes relating to continuous integration in github*. Em *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, páginas 805–816, 2015. 10, 11
- [32] Zhao, Yangyang, Alexander Serebrenik, Yuming Zhou, Vladimir Filkov e Bogdan Vasilescu: *The impact of continuous integration on other software development practices: a large-scale empirical study*. Em *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, páginas 60–71. IEEE, 2017. 11, 28, 30, 31, 36, 69
- [33] Wessel, Mairieli, Alexander Serebrenik, Igor Wiese, Igor Steinmacher e Marco A Gerosa: *Effects of adopting code review bots on pull requests to oss projects*. Em *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, páginas 1–11. IEEE, 2020. 11, 28, 29, 31, 69
- [34] Ortu, Marco, Giuseppe Destefanis, Steve Counsell, Stephen Swift, Roberto Tonelli e Michele Marchesi: *How diverse is your team? investigating gender and nationality diversity in github teams*. *Journal of Software Engineering Research and Development*, 5(1):1–18, 2017. 11
- [35] Meyer, Andre N, Laura E Barton, Gail C Murphy, Thomas Zimmermann e Thomas Fritz: *The work life of developers: Activities, switches and perceived productivity*. *IEEE Transactions on Software Engineering*, 43(12):1178–1193, 2017. 12
- [36] Beller, Moritz, Vince Orgovan, Spencer Buja e Thomas Zimmermann: *Mind the gap: On the relationship between automatically measured and self-reported productivity*. *IEEE Software*, 2020. 12
- [37] Gousios, Georgios, Andy Zaidman, Margaret Anne Storey e Arie Van Deursen: *Work practices and challenges in pull-based development: The integrator’s perspective*. Em *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 1, páginas 358–368. IEEE, 2015. 12, 39
- [38] Steinmacher, Igor, Marco Aurelio Graciotto Silva, Marco Aurelio Gerosa e David F Redmiles: *A systematic literature review on the barriers faced by newcomers to open source software projects*. *Information and Software Technology*, 59:67–85, 2015. 12
- [39] Calefato, Fabio, Marco Aurelio Gerosa, Giuseppe Iaffaldano, Filippo Lanubile e Igor Steinmacher: *Will you come back to contribute? investigating the inactivity of oss core developers in github*. arXiv preprint arXiv:2103.04656, 2021. 12, 16, 27, 50, 62
- [40] Silva, Jefferson, Igor Wiese, Daniel M German, Christoph Treude, Marco Aurélio Gerosa e Igor Steinmacher: *A theory of the engagement in open source projects via summer of code programs*. Em *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, páginas 421–431, 2020. 12

- [41] Miller, Courtney, David Gray Widder, Christian Kästner e Bogdan Vasilescu: *Why do people give up flossing? a study of contributor disengagement in open source*. Em *IFIP International Conference on Open Source Systems*, páginas 116–129. Springer, 2019. 12
- [42] Bao, Lingfeng, Xin Xia, David Lo e Gail C Murphy: *A large scale study of long-time contributor prediction for github projects*. *IEEE Transactions on Software Engineering*, 2019. 12, 24
- [43] Pinto, Gustavo, Igor Steinmacher e Marco Aurélio Gerosa: *More common than you think: An in-depth study of casual contributors*. Em *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, volume 1, páginas 112–123. IEEE, 2016. 12, 47
- [44] Gousios, Georgios, Martin Pinzger e Arie van Deursen: *An exploratory study of the pull-based software development model*. Em *Proceedings of the 36th International Conference on Software Engineering*, páginas 345–355, 2014. 12
- [45] Joblin, Mitchell, Sven Apel, Claus Hunsen e Wolfgang Mauerer: *Classifying developers into core and peripheral: An empirical study on count and network metrics*. Em *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, páginas 164–174. IEEE, 2017. 13
- [46] Mockus, Audris, Roy T Fielding e James D Herbsleb: *Two case studies of open source software development: Apache and mozilla*. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 11(3):309–346, 2002. 13
- [47] Avelino, Guilherme, Leonardo Passos, Andre Hora e Marco Tulio Valente: *A novel approach for estimating truck factors*. Em *2016 IEEE 24th International Conference on Program Comprehension (ICPC)*, páginas 1–10. IEEE, 2016. 13, 14, 26, 48, 51, 69
- [48] Guzman, Emitza, David Azócar e Yang Li: *Sentiment analysis of commit comments in github: an empirical study*. Em *Proceedings of the 11th working conference on mining software repositories*, páginas 352–355, 2014. 15, 58
- [49] Pletea, Daniel, Bogdan Vasilescu e Alexander Serebrenik: *Security and emotion: sentiment analysis of security discussions on github*. Em *Proceedings of the 11th working conference on mining software repositories*, páginas 348–351, 2014. 15
- [50] Sinha, Vinayak, Alina Lazar e Bonita Sharif: *Analyzing developer sentiment in commit logs*. Em *Proceedings of the 13th international conference on mining software repositories*, páginas 520–523, 2016. 15
- [51] Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai e Arvid Kappas: *Sentiment strength detection in short informal text*. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010. 15
- [52] Loper, Edward e Steven Bird: *Nltk: The natural language toolkit*. arXiv preprint cs/0205028, 2002. 15

- [53] Islam, Md Rakibul e Minhaz F Zibran: *Leveraging automated sentiment analysis in software engineering*. Em *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, páginas 203–214. IEEE, 2017. 16
- [54] Islam, Md Rakibul e Minhaz F Zibran: *Deva: sensing emotions in the valence arousal space in software engineering text*. Em *Proceedings of the 33rd annual ACM symposium on applied computing*, páginas 1536–1543, 2018. 16
- [55] Calefato, Fabio, Filippo Lanubile, Federico Maiorano e Nicole Novielli: *Sentiment polarity detection for software development*. *Empirical Software Engineering*, 23(3):1352–1382, 2018. 16, 28, 58
- [56] Ahmed, Toufique, Amiangshu Bosu, Anindya Iqbal e Shahram Rahimi: *Senticr: a customized sentiment analysis tool for code review interactions*. Em *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, páginas 106–111. IEEE, 2017. 16
- [57] Novielli, Nicole, Daniela Girardi e Filippo Lanubile: *A benchmark study on sentiment analysis for software engineering research*. Em *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*, páginas 364–375. IEEE, 2018. 16
- [58] Novielli, Nicole, Fabio Calefato, Davide Dongiovanni, Daniela Girardi e Filippo Lanubile: *Can we use se-specific sentiment analysis tools in a cross-platform setting?* Em *Proceedings of the 17th International Conference on Mining Software Repositories*, páginas 158–168, 2020. 16, 28, 58, 69
- [59] Ford, Denae, Margaret Anne Storey, Thomas Zimmermann, Christian Bird, Sonia Jaffe, Chandra Maddila, Jenna L. Butler, Brian Houck e Nachiappan Nagappan: *A tale of two cities: Software developers working from home during the covid-19 pandemic*, 2020. 17, 63, 67, 68
- [60] Lal, Banita, Yogesh K Dwivedi e Markus Haag: *Working from home during covid-19: How do we ‘do’ social interaction at a distance?* Em *International Working Conference on Transfer and Diffusion of IT*, páginas 320–328. Springer, 2020. 17
- [61] Xiong, Ziyu, Pin Li, Hanjia Lyu e Jiebo Luo: *From gen z, millennials, to baby-boomers: Portraits of working from home during the covid-19 pandemic*. arXiv preprint arXiv:2101.06762, 2021. 17
- [62] Garro-Abarca, Victor, Pedro Palos-Sanchez e Mariano Aguayo-Camacho: *Virtual teams in times of pandemic: Factors that influence performance*. *Frontiers in Psychology*, 12:232, 2021. 18
- [63] Georgiou, Konstantinos, Nikolaos Mittas, Lefteris Angelis e Alexander Chatzigeorgiou: *A study of knowledge sharing related to covid-19 pandemic in stack overflow*, 2020. 18
- [64] Rahman, Akond e Effat Farhana: *An empirical study of bugs in covid19 software projects*. *Journal of Software Engineering*, 9:3, 2021. 19

- [65] Oliveira, Pedro Almir Martins de, Pedro de Alcântara dos Santos Neto, Gleison Silva, Irwayne Ibiapina, Werney Lira e Rossana Maria de Castro Andrade: *Software development during covid-19 pandemic: an analysis of stack overflow and github*. arXiv preprint arXiv:2103.05494, 2021. 19
- [66] Butt, Shariq Aziz, Sanjay Misra, Muhammad Waqas Anjum e Syed Areeb Hassan: *Agile project development issues during covid-19*. Em *International Conference on Lean and Agile Software Development*, páginas 59–70. Springer, 2021. 19
- [67] Marek, Krzysztof, Ewelina Wińska e Włodzimierz Dabrowski: *The state of agile software development teams during the covid-19 pandemic*. Em *International Conference on Lean and Agile Software Development*, páginas 24–39. Springer, 2021. 19
- [68] Klotzman, Vanessa, Farima Farmahinifarahani e Cristina Lopes: *Public software development activity during the pandemic*. Em *Proceedings of the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, páginas 1–12, 2021. 20
- [69] Gousios, Georgios: *The ghtorrent dataset and tool suite*. Em *Proceedings of the 10th Working Conference on Mining Software Repositories, MSR '13*, páginas 233–236, Piscataway, NJ, USA, 2013. IEEE Press, ISBN 978-1-4673-2936-1. <http://dl.acm.org/citation.cfm?id=2487085.2487132>. 21
- [70] Gyimesi, Péter, Béla Vancsics, Andrea Stocco, Davood Mazinianian, Arpád Beszédes, Rudolf Ferenc e Ali Mesbah: *Bugsjs: A benchmark of javascript bugs*. Em *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*, páginas 90–101. IEEE, 2019. 24
- [71] Pino, Francisco Alberto: *A questão da não normalidade: Uma revisão*. *Revista de economia agrícola*, 61(2):17–33, 2014. 28
- [72] Cliff, Norman: *Dominance statistics: Ordinal analyses to answer ordinal questions*. *Psychological bulletin*, 114(3):494, 1993. 29
- [73] Gałecki, Andrzej e Tomasz Burzykowski: *Linear mixed-effects model*. Em *Linear Mixed-Effects Models Using R*, páginas 245–273. Springer, 2013. 31
- [74] Cassee, Nathan, Bogdan Vasilescu e Alexander Serebrenik: *The silent helper: the impact of continuous integration on code reviews*. Em *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, páginas 423–434. IEEE, 2020. 31
- [75] Wilks, Daniel S: *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011. 34
- [76] Romano, Jeanine, Jeffrey D Kromrey, Jesse Coraggio e Jeff Skowronek: *Appropriate statistics for ordinal level data: Should we really be using t-test and cohen'sd for evaluating group differences on the nsse and other surveys*. Em *annual meeting of the Florida Association of Institutional Research*, volume 177, 2006. 34

- [77] Bernardo, João Helis, Daniel Alencar da Costa e Uirá Kulesza: *Studying the impact of adopting continuous integration on the delivery time of pull requests*. Em *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*, páginas 131–141. IEEE, 2018. 39
- [78] Fronchetti, Felipe, Igor Wiese, Gustavo Pinto e Igor Steinmacher: *What attracts newcomers to onboard on oss projects? tl; dr: Popularity*. Em *IFIP International Conference on Open Source Systems*, páginas 91–103. Springer, 2019. 45
- [79] Steinmacher, Igor, Tayana Conte, Marco Aurélio Gerosa e David Redmiles: *Social barriers faced by newcomers placing their first contribution in open source software projects*. Em *Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing*, páginas 1379–1392, 2015. 45
- [80] Lee, Amanda, Jeffrey C Carver e Amiangshu Bosu: *Understanding the impressions, motivations, and barriers of one time code contributors to floss projects: a survey*. Em *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, páginas 187–197. IEEE, 2017. 46