MASTER

Effect of the introduction of code of conduct in collaborative development

How does the introduction of codes of conduct influence women's involvement in Open Source Software?

Enache, Bogdan

Award date:
2021

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computer Science
Software Engineering and Technology Research group

Effect of the introduction of code of conduct in collaborative development

How does the introduction of codes of conduct influence women's involvement in Open Source Software?

Author:
Bogdan Enache
1035066

Supervisor:
Dr. Eleni Constantinou (TU/e)

Co-supervisor:
Prof. dr. Alexander Serebrenik (TU/e)

Committee:
Dr. Irina Kostitsyna (TU/e)

Version 1.0

Eindhoven, July 19, 2021

Abstract

Codes of conduct represent a sign of a welcoming community, aimed at protecting people, especially minorities, from harassment by enforcing a behavioural set of rules inside the OSS community. Furthermore, code of conduct stresses the desire of diversity in OSS communities [52]. In this way, authors of code of conduct hope to attract people to join OSS communities [52].

In OSS, women represent a minority which is vulnerable when it comes to harassment in OSS communities [1,3,17,23,55,59]. Studies have shown unaccepted behaviour towards women such as sexualizing and offensive talk, in OSS communities [33]. Vertical and horizontal segregation with respect to women is a reality [10]. Some OSS contributors are even strongly opposed to the inclusion of women in OSS communities [27]. In a recent study by Singh et al. [48], women perceive code of conduct as a tangible solution for removing offensive behaviour towards women. Furthermore, women's perception is that the adoption of code of conduct in an OSS community, makes them engage more and feel more comfortable in the OSS community, giving them a sense of belonging in the OSS community [48].

To determine whether code of conduct has indeed any influence on women in OSS communities, this study explores how women's involvement is influenced after a code of conduct is adopted in OSS projects. A total of 418 GitHub projects, that have adopted a code of conduct, were analysed from a quantitative point of view, by applying statistical models such as Regression Discontinuity Design, Survival Analysis, Log-Rank test and Cox regression. The conclusion from the statistical analysis was that code of conduct has a momentary, positive impact, on women's involvement, right after the adoption of code of conduct, but no lasting impact on women's involvement in OSS projects.

Preface

Firstly, I want to thank my supervisors, Eleni Constantinou and Alexander Serebrenik, for guiding and helping me throughout the entirety of my thesis. I would also like to thank Irina Kostitsyna for joining my committee as a third committee member, and Elian Carsenat for helping me in using the Namsor tool for identifying the gender of GitHub users.

I would also like to thank my family and friends for supporting me throughout the entire years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The concept of code of conduct is not new. It has been used for many years by companies as a way to establish a company culture and boost productivity of their employees [14]. Codes of conduct are especially useful when companies have a large number of employees or when their employees come from different countries, have different backgrounds or experiences; since these are known to create conflicts and tension amongst co-workers [14]. For companies, code of conduct represents a set of behavioural rules that the individuals from an organization need to adhere to, if they want to be part of it. People who violate the rules are sanctioned. In this way, codes of conduct aim at creating a safe environment, in which everyone can contribute in peace, without facing any sort of discrimination or offensive behaviour.

Open source software projects are, based on the definition made by Open Source[1], software projects which have the source code made publicly available to be used, modified and shared by anyone. Thanks to software technologies, such as version control systems and chat systems, people from different countries, backgrounds and experiences can work together in creating, maintaining and improving open source software projects [52]. This can lead to some OSS projects having very diverse teams, increasing the risk of offensive behaviour [52]. An example of such behaviour took place in the OpalRB, an OSS project, where one of the many core maintainers made a transphobic post on Twitter[2]. Developers, project members and people not affiliated with the project had mixed opinions, some of them demanding no consequences while others calling for his exclusion from the project[3]. The consequence of this discussion is that the Opal community has decided to adopt a code of conduct in order to prevent offensive behaviour from ever happening again in their community.

Transphobic behaviour is not the only type of offensive behaviour that happens in OSS communities. Another type of offensive behaviour, out of many others, is related to gender. Studies have shown that women face misogyny, sexism, harassment and discrimination in OSS communities [1, 3, 17, 23, 33, 55, 59]. Analysing the perception of women about code of conduct in OSS projects, a recent study, by Singh et al. [48], has shown that the adoption of code of conducts in OSS projects might prevent such offensive behaviour towards women, as well as might motivate women to engage more in OSS communities which have adopted a code of conduct. Furthermore, women reported that gender-related incidents might represent a cause why women leave OSS projects sooner [55]. Based on the aforementioned studies, codes of conduct might have an impact on women's involvement in OSS projects.

The importance that code of conduct might have in OSS communities, has been recognized even by GitHub[4], the largest online platform for hosting OSS projects[5], which now asks users, when creating a new software project, whether they would like to add a code of conduct template in their project. The added code of conduct is in the form of a text file which is stored in the project. GitHub claims that the role of the code

[1] https://opensource.com/resources/what-open-source
[2] https://twitter.com/krainboltgreene/status/611569515315507200
[3] https://github.com/opal/opal/issues/941
[4] https://github.com/
[5] https://octoverse.github.com/

of conduct is to signal an inclusive project, welcoming everyone to contribute and outline procedures for handling abuse[6].

Based on a study from 2017 by Tourani et al. [52], there are different codes of conduct, created specifically for OSS projects, made by Ubuntu[7], Contributor Covenant[8], Django[9], Python[10], Mozilla[11] and so on. Despite having codes of conduct written by different communities, such as Python, or individuals, such as Contributor Covenant, the study by Tourani et al. [52] has shown that all of them represent a sign of a welcoming community, aimed at protecting people, especially minorities, from harassment. Furthermore, they all contain a set of behavioural rules, listing what is acceptable and what not. There are also people, in an OSS project, which are selected for enforcing the guidelines. People who violate the rules might face some consequences, based on how serious the offence was. For serious offences, the consequences might lead to the removal of the person who violated the rules.

While in recent years, researchers have become more interested in codes of conduct in OSS projects, analyzing its popularity and what it promises (Tourani et al. [52]; 2017), the perception that women have on code of conduct (Singh et al. [48]; 2021), as well as the impact that code of conduct has on women's presence in OSS projects (Neill Robson [43]; 2018), there is still few research about code of conduct.

This thesis proposes to determine, based on a quantitative analysis, the impact that code of conduct has on women's involvement in OSS. The reason for choosing women as the focus group is based on the offensive behaviour they might face in OSS projects. Another reason for choosing women as the focus group is that having a code of conduct in an OSS project, might offer women a sense of belonging in the OSS community and might motivate women to engage more in OSS projects [48]. Another argument is that women in OSS are relatively well studied as opposed to other minorities. The only quantitative research, for this topic, was done by Neill Robson [43], but their analysis is focused only on 500 projects having the largest number of pull requests. In comparison to Robson's work, this thesis aims to take into account all projects which have adopted a code of conduct on GitHub, for having a better understanding of the impact that code of conduct has on women's involvement, as well as taking into account for confounds, such as the passing of time, which might influence results. The focus of this thesis is the following one:

How does the adoption of code of conduct influence women's involvement in OSS projects?

The focus of this thesis is on determining whether code of conduct has influences on women's involvement. One possible impact that code of conduct might have on women's involvement, is related with their presence in the OSS projects. This might result in a higher proportion of women in OSS projects, in the period after code of conduct was adopted, in comparison to the period before code of conduct was adopted. Another possible impact that code of conduct might have on women's involvement, is the time women remain contributors in a project.

For the purpose of this thesis, only the OSS projects stored on GitHub are taken into account, since it is the largest online platform for hosting OSS projects.

1.1 Research questions

The focus of this thesis is to determine how the code of conduct influences women's involvement in OSS projects. Since this objective is rather vague / broad, it needs to be split into more concrete goals based on which more specific research questions can be formulated.

Since the focus is on involvement of women, we consider two methods of measuring the involvement: how many women have been involved in OSS projects, and for how long. The reason for focusing on measuring involvement, and not on measuring contribution, is that the current approaches for measuring contribution,

[6] https://docs.github.com/en/communities/setting-up-your-project-for-healthy-contributions/adding-a-code-of-conduct-to-your-project

[7] https://ubuntu.com/community/code-of-conduct

[8] https://contributor-covenant.org/version/2/0/code_of_conduct/

[9] https://djangoproject.com/conduct/

[10] https://python.org/psf/conduct/

[11] https://mozilla.org/en-US/about/governance/policies/participation/

such as the number of commits or number of pull requests, number of lines of code, cannot do justice to the extent of a contribution that a person had in an OSS project [5, 36, 56, 62].

A way to measure how many women have been involved in OSS projects, is to identify how many women have contributed to a project, out of the total people who have contributed in the respective project. In essence, out of the total contributions for each project, what percentage of it is done by women.

Measuring how long a woman was involved in a project can be computed by taking into account the time passed between the date of the first contribution they made, and the date of the last contribution they made in the respective project.

Based on everything mentioned until now, the following two research questions can be formulated:

- **RQ1:** Do projects have a (higher) increase in the proportion of women who contribute in the project, in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?
- **RQ2:** Do women remain contributors over a longer duration of time in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?

1.2 Report structure

The report has the following structure: Chapter 2 presents the related research of codes of conduct. Chapter 3 presents how the data was collected and processed as well as how the data was analyzed to answer **RQ1** and **RQ2**. Chapter 4 presents the results and answers for **RQ1** and **RQ2**. Chapter 5 discusses the results based on which a conclusion is drawn and future work is presented.

Chapter 2

Related Work

In this chapter, the relevant literature about code of conduct in OSS projects is discussed as well as the offensive behaviour towards women in OSS.

2.1 Code of conduct

At the time of writing this thesis, there are few research papers which analyze the impact of code of conduct in OSS projects.

The work done by Tourani et al. [52], in which the scope, role and influence of codes of conduct are analyzed, represents the first paper focused solely on code of conduct. In this paper, thirteen codes of conduct, from the ToDo group[1], were investigated to determine their popularity, based on the number of GitHub projects which have adopted it. The conclusion was that Contributor Covenant was the most popular code of conduct and that codes of conduct are used by more than 500 OSS projects. Furthermore, the thirteen codes of conduct were manually studied from five dimensions: purpose, honorable behaviour, enforcement and scope. It was concluded that all of them outline the same set of values and expectations on how people should behave. The scope of code of conduct is to represent a sign of a welcoming community, aimed at protecting people, especially minorities, from harassment. To understand how code of conducts are used in OSS projects, interviews with leaders of OSS projects were held. The aim of the interviews was to understand the motivation, content, evolution and enforcement policy of codes of conduct. The results from the interviews were that leaders of OSS projects adopted codes of conduct, as an attempt to avoid offensive behaviour. Code of conduct contains a set of behavioural rules and for enforcing the code of conduct, two approaches are used in OSS projects: either the code of conduct is enforced by the OSS project maintainers, or every member from the community signs to follow the code of conduct rules. The enforcement policy was done by selected members from the OSS project. After manually analysing the content of code of conduct, as well as performing interviews with OSS project maintainers, the conclusion, of the paper done by Tourani et al. [52], is that code of conduct promotes an inclusive and friendly environment, targeted especially to the people who face harassment such as minorities. In comparison to the work done by Tourani et al. [52], this thesis focuses on analyzing the impact that code of conduct has on a minority group of OSS projects, more specifically, women in OSS projects. The work done by Tourani et al. [52], does not analyze the impact, if any, that code of conduct has on women, or any other minority, instead, it focuses on what OSS maintainers strive to achieve by adopting code of conduct, without determining how effective they are in OSS projects.

Another quantitative and qualitative research was done by Qiu et al. [41], in which it was analyzed what signs do people look for when deciding on which OSS project to contribute to. For determining what signs contributors look for, when deciding on which OSS projects to contribute to, interviews with OSS contributors were held. Based on the results of the interviews, a set of signals was created. The set of signals that contributors look for when deciding on which OSS projects to contribute, were tested by mining GitHub repositories and applying statistical models. One of the signals included in the research, is the presence of

[1] https://todogroup.org/

code of conduct, for which it was concluded that code of conduct has little to no impact to people when they are deciding whether to contribute or not to an OSS project.

A research that is based solely on quantitative analysis, is the one done by Robson [43], which investigates whether the code of conduct has a significant change in the proportion of women who contribute in projects with code of conduct, in comparison to projects which do not have a code of conduct. In their work, 500 GitHub projects, by number of pull requests, were considered. The proportion of women was computed by splitting the history of each project into 24 hours time periods. For determining whether the code of conduct has any influence on women's presence, the projects which did not have a code of conduct were compared, using statistical tests, to the projects which have a code of conduct. The result of the statistical tests, applied in Robson's work [43], was that there is no statistical significant difference in women's presence between the projects which did not have a code of conduct and to the projects which have a code of conduct.

In comparison to Robson's work [43], this thesis aims at solving several threats from their work. The first potential threat is the choice of projects, since only the top 500 GitHub repositories, by number of pull requests, are considered in their work. Their choice of projects might impact the generalization of the results. To improve the generalization of the results, our research aims at taking into account all the GitHub projects which have adopted a code of conduct, regardless of their popularity or number of pull requests. A second potential threat is related to relying only on commits as a way to identify contributors. As shown in the work of Young et al. [62], presented in Section 3.2, relying only on commits represents an invalid method of identifying contributors, since people's contribution is more comprehensive than just commits. To get a more accurate representation of the projects' communities, contributors are identified using three of the contributorship models defined by Young et al. [62], presented in Section 3.2. A third potential threat, for Robson's work, is related with the statistical models used in their work. In their work, Wilcoxon Signed Rank statistical test is used to determine the impact that code of conduct has on proportion of women in OSS. Not taking into account for the presence of confounds, might influence the impact that code of conduct has on proportion of women. Furthermore, to evaluate the effect of a treatment or change (e.g., a new drug on a disease, a new rule on a community) a longitudinal study is suited [9]. Since the impact of adopting a code of conduct might span across a long period of time (i.e. the community does not change over night simply because a code of conduct was adopted, instead, such changes need time to be adopted), a statistical model, which takes into account changes over time, might represent a better option for determining whether code of conduct has an impact on women's proportion. To address this threat, in this thesis, a mixed-effect model, which takes into account for confounds [4], is used, as well as a statistical model, based on Regression Discontinuity Design (RDD), which takes into account changes over time.

From a qualitative research, the paper by Singh et al. [48], offers a better understanding on how women perceive the presence of code of conduct in OSS projects. In finding an answer, the research was focused on analyzing the messages about code of conduct from women-focused forums. The women-focused forums are also seen as a safe space for women. A total of 10,689 messages from 1344 participants were analysed. The conclusion was that women face sexism and misogyny in OSS communities. The results found, in the paper by Singh et al. [48], show that the hostility, in OSS projects, towards women, stops women from contributing in OSS. **The conclusion was that most of them feel that the presence of a code of conduct might represent an answer to prevent offensive behaviour. Women also stated that the presence of code of conduct motivates them to engage more in the project, since it gives them a sense of belonging. Women also reported positive experiences in projects which have a code of conduct.** The conclusions from this paper are matching also with the intuition that Neill Robson had about the code of conduct, in his paper [43].

2.2 Women in open source software

In this section, the related work about the offensive behaviour and discrimination faced by women in the software field, is presented.

A study by Nafus [33], has shown that the OSS community is not open to receiving women as contributors. After interviewing women in key positions in OSS communities, it was concluded there are many women who are facing constant discrimination [33], some even reporting that "men monopolize code authorship and simultaneously de-legitimize the kinds of social ties necessary to build mechanisms for women's inclusion"

[33]. Furthermore, no matter the knowledge and expertise, both horizontal and vertical gender segregation with respect to women are a reality, only 2.3% of women are in top positions in OSS [10]. Studies have also shown that, in the presence of discrimination, women are more likely to report the “impostor syndrome” [56].

The offensive behaviour towards women is further amplified when women stated that the software field is seen as a male-dominated activity in which only men are capable of performing well [59]. Because of this, when technical decisions have to be taken, members of the OSS community are more inclined to trust men’s point of view in the detriment of women’s, leading to gender biases when a technical decision needs to be taken [59].

Some OSS contributors are strongly opposed to the inclusion of women [27], making it difficult to be accepted in OSS as a woman [27]. For women who want to fit or contribute in an OSS community, they feel the constant pressure to masculinize themselves [1]. One possible solution for being accepted in an OSS project and avoid offensive behaviour or discrimination, is for women to use OSS accounts with male names [32, 55], since gender is the second most visible attribute in OSS [55]. People are also aware of the gender of most of the other people within the team [56]. Since the gender of contributors is known, and taking into account the fact that some OSS contributors do not accept women or that contributors are more inclined to trust men’s point of view, it might might have an impact on women’s involvement in OSS projects.

OSS is a field predominated by men and gender biases that prevent women from joining the software industry [23, 59]. Ratio of male is higher than that of females in OSS [57]. Women are a minority in OSS, their contribution in OSS count for 5% of total contributions in OSS [10], despite studies showing that higher gender diversity in OSS leads to higher productivity [56]. When interviewing women, they associated the high percentage of males in OSS projects with a negative impact on the communication styles in OSS, making women feel not welcomed and uncomfortable [3, 33].

For both men and women, the motivations to contribute to OSS are similar, they both want to learn new programming and management skills, but the way they engage is different [3]. Firstly, men choose an OSS project to contribute on, mostly on popularity, because they feel that it is their duty and see it more as a job, while women choose a project because they have a real interest in it or the community around it [3].Women stated that “for men it’s more their job to contribute to OSS, but women want to do it because they find it exciting” [3] or “It’s more difficult for women to stick around also, the top reason is that it’s not their job – they’re not being paid to do it” [3]. Thus, an intuition might be that women might join a project because it is exciting and women might pay more attention towards how they feel being part of the OSS project, in comparison to men.

Gender-related incidents are also a cause why women leave projects sooner [55]. As a consequence to the offensive behaviour that women face in OSS projects, women might engage less in OSS [3,23] and contribute to fewer projects because of gender biases [23] and sexist behaviour [17]. The lack of peer parity (seeing other women in the OSS community) is another reason why women are less likely to join OSS projects, but more likely to engage less or stop contributing [16]. A systematic literature review, in which all the research papers related with “women in OSS projects” were analyzed (24 papers in total), concluded that women engage less due to gender bias and offensive behaviour [11].

The gender bias and the bad behaviour towards women have a negative impact on productivity. Studies have shown that women’s involvement proves to be positive in OSS [56]. A higher gender diversity in OSS leads to higher productivity [56]. Furthermore, based on a recent study, over 30% of men and 60% of women think that gender diversity has a positive impact to software development [6]. Determining if the adoption of code of conduct has a positive impact on proportion of women and on the time women remain contributors in OSS projects, might be also beneficial for OSS project maintainers, since it might boost the productivity in their OSS projects.

Chapter 3

Methodology

This chapter describes what research strategy is used for the quantitative analysis of this thesis, what data is needed for the quantitative analysis of this thesis, as well as how it is extracted from online platforms which store OSS projects. The research strategy is described in Section 3.1. The definition of contribution, used for identifying contributors in OSS projects, is presented in Section 3.2. The data needed for answering **RQ1** and **RQ2** is described in Section 3.3. The platform from which data is collected is described in Section 3.4. Selecting the GitHub repositories which have a code of conduct is discussed in Section 3.5. The identified repositories are filtered based on certain criteria, as presented in Section 3.6. How the contribution data is extracted from the remaining repositories and how the contributors are identified is discussed in Section 3.7. Identifying the gender of contributors is presented in Section 3.8. The hypotheses needed to answer **RQ1** and **RQ2** are presented in Section 3.9, the statistical techniques are presented in Sections 3.10 and 3.11 and the statistical models are described in Section 3.12.

3.1 Research strategy and method

Based on **RQ1** and **RQ2**, described in Section 1.1 a research strategy is selected, based on the research strategies defined by Stol et al. [49]. According to the selected strategy, a research method is chosen.

The research strategy most suited for such research questions is a *sample study* [49], since it allows to analyze and identify correlations between characteristics in a population. For this strategy, only a sample of the population is taken and analysed, as suggested by the name. This strategy has at its core, statistics, which are useful for analyzing the impact that code of conduct has on women's involvement. The three main research methods, adapted to the purpose of this thesis, are:

- **Repository mining:** extracting data from repositories that are stored on GitHub, in order to determine the impact that code of conduct has on women's involvement;
- **Interviews:** talk with people who contribute in the repositories that are stored on GitHub, in order to determine the impact that code of conduct has on women's involvement;
- **Surveys:** send a survey to people who contribute in the repositories that are stored on GitHub, in order to determine the impact that code of conduct has on women's involvement

Because of time constraints, this research is solely based on repository mining. Data is extracted from GitHub, processed and analysed using statistics models in order to answer the research questions and derive a conclusion.

Before doing any repository mining and applying any statistical models, it is important to decide how, and for what, we want to use this data in order to answer **RQ1** and **RQ2**. To determine whether code of conduct has any impact on the proportion of women, or time women remain contributors in a project, different approaches are possible:

- **Approach 1:** analyze a group of women which contribute in both OSS projects with and without code of conduct, and determine if the proportion of women is higher and if women contributed for longer periods, in OSS projects which have a code of conduct in comparison to the ones which do not. The problem with this approach is that, there is too much variability across projects, since projects differ vastly from one to another, thus comparing data from projects with code of conduct, to data from projects without code of conduct might lead to wrong results and conclusions. Furthermore, this approach does not tell anything about how the code of conduct changed the OSS projects after its adoption, in comparison to before;

- **Approach 2:** Take measurements in projects which have a code of conduct and in projects which have never had a code of conduct and then compare the data to determine if the proportion of women is higher and if women contributed for longer periods, in OSS projects which have a code of conduct in comparison to the ones which do not. This approach has the same problems as the one mentioned in **Approach 1**.

- **Approach 3:** Measure OSS projects, which were created without a code of conduct, but which have adopted a code of conduct later in their history. These projects have a part of their history without a code of conduct, which is referenced throughout the entire thesis as the **pre-conduct** period, and the rest of their history with a code of conduct, which is referenced as the **post-conduct** period. For this approach, the proportion of women who contributed in the pre-conduct period, and in the post-conduct period, of the same projects, are compared to determine if the proportion of women is higher and if women contributed for longer periods, in the post-conduct period, in comparison to the pre-conduct period.

Approach 3 was selected, since determining the impact that code of conduct has on proportion of women in OSS projects, and on the time women remain contributors in OSS project, is more robust than the one mentioned in **Approach 1** and **Approach 2** since the measurements are taken in the same projects, removing the variability across projects. Furthermore, this approach has been used in other papers as well, such as the one by Zhao et al. [63], in which the impact that code of conduct had on software development practices. Another paper which used this approach was the one by Trockman et al. [53], in which the impact of badges in the npm ecosystem was analyzed.

3.2 Defining contribution

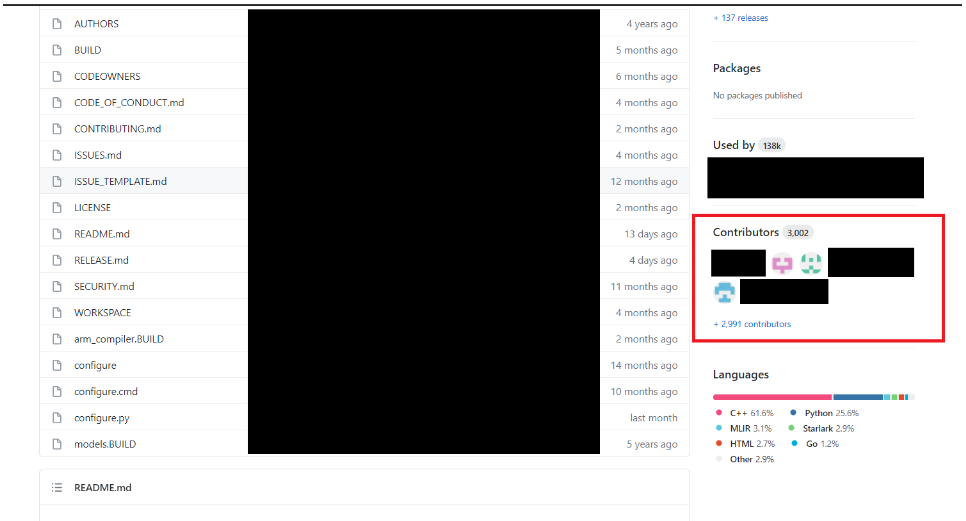
Answering **RQ1** and **RQ2** requires identifying women who are contributors in OSS projects. A study by Vasilescu et al. [56], in which contributors were asked what is a contributor, the conclusion of the survey was that a contributor is any person who makes a contribution. The question that one might ask is: What does contribution mean? What is the best metric, on GitHub, to reflect contribution? For this, the literature had a recent shift. Commits are seen as *"the most encompassing form of coding contribution to a GitHub project"* [56] and represent a good way of measuring the productivity [12,56]. There is even an entire research on commits as a metric to reflect contribution [5]. Instead of relying solely on commits, for this research, it was decided to use the contribution defined in the paper by Young et al. [62]. What is different about the paper by Young et al., is that it takes into account the contributors that are identified based on GitHub interactions (such as making commits, pull-requests, comments, replies, issues, as well as the contributors who cannot be identified based on GitHub interactions (e.g.,: people who are in charge of organizing events for the OSS project or people who deal with marketing, finance and other sectors that are not related to GitHub interactions). To take into account both categories of contributors, the paper by Young et al. defines four models of contributorship. This is especially useful since non-code contribution is hard to track, since GitHub is focused mostly on code contribution. The four models of contribution are the following:

- **Model 1 of contributorship:** Contributors identified based on the Contributors list shown by GitHub, as can be seen in Figure 3.1. Problem with this method is that it only shows the top 100 contributors of the project (i.e. the ones with the most commits);

- **Model 2 of contributorship:** Contributors identified based on the interactions with the repository (commits, pull requests, issues, discussions, comments, replies, reactions and so on). In essence, everything that can be extracted from the GitHub repository. Tools have been built for identifying such contributors[1];

- **Model 3 of contributorship:** Contributors identified based on taxonomies. Here they have identified the contributors based on the ".all-contributors" file from the All Contributors model[2]. The All Contributors model is meant to give credit to everyone who has contributed in a GitHub repository, both code and non-code contribution. The ".all-contributors" file is of format JSON, and contains a list with all the people who have ever contributed in the GitHub repository. This file is stored inside the GitHub repository, as can be seen in Figures 3.2 and 3.3;

- **Model 4 of contributorship:** Contributors identified ad-hoc. These are the contributors that can be identified by parsing non-standardized data sources (e.g., by parsing text files such as "contributors" and "authors"). Its purpose is very similar to the All Contributors model, the only difference being that the files do not have a standardised format. Instead of a JSON format, they usually contain a bullet point list, or table with the names of the contributors, as can be seen in Figure 3.4. Furthermore, they might contain other text besides the list (such as a thank you message for example). The main problem with this model is how hard is to extract automatically the data from these text files.



AUTHORS	[redacted]	4 years ago
BUILD		5 months ago
CODEOWNERS		6 months ago
CODE_OF_CONDUCT.md		4 months ago
CONTRIBUTING.md		2 months ago
ISSUES.md		4 months ago
ISSUE_TEMPLATE.md		12 months ago
LICENSE		2 months ago
README.md		13 days ago
RELEASE.md		4 days ago
SECURITY.md		11 months ago
WORKSPACE		4 months ago
arm_compiler.BUILD		2 months ago
configure		14 months ago
configure.cmd		10 months ago
configure.py		last month
models.BUILD		5 years ago

README.md

+ 137 releases

Packages

No packages published

Used by 138k

[redacted]

Contributors 3,002

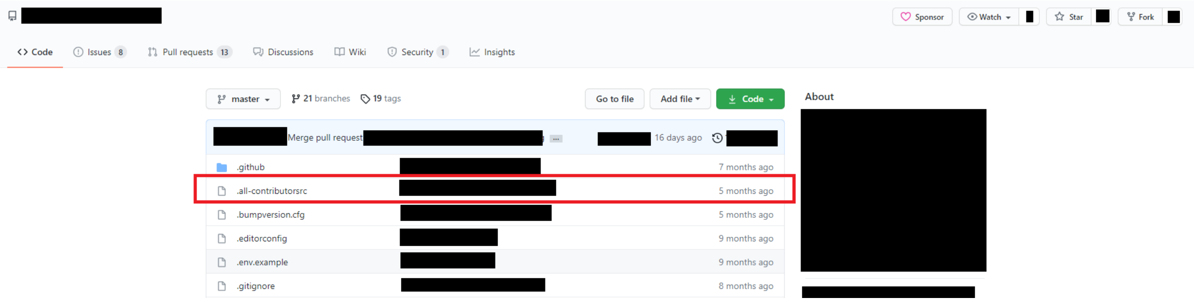
[redacted] [redacted] [redacted]

+ 2,991 contributors

Languages

● C++ 61.6% ● Python 25.6%
● MLIR 3.1% ● Starlark 2.9%
● HTML 2.7% ● Go 1.2%
● Other 2.9%

Figure 3.1: In red, the contributors list as shown by GitHub. Sensitive information has been masked.



[redacted]

Sponsor | Watch [redacted] | Star [redacted] | Fork [redacted]

Code | Issues 8 | Pull requests 13 | Discussions | Wiki | Security 1 | Insights

master | 21 branches | 19 tags | Go to file | Add file | Code

About

[redacted]

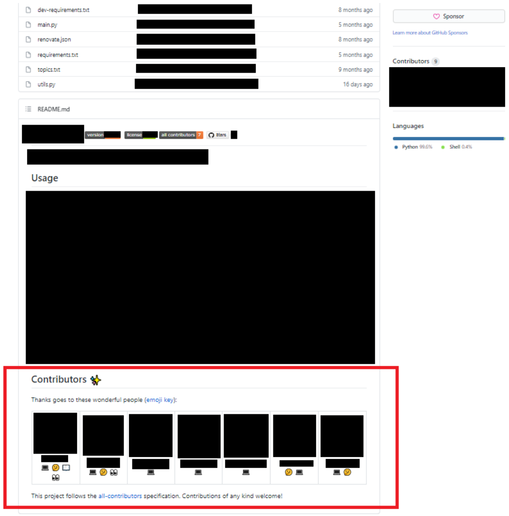
[redacted] Merge pull request [redacted] ... [redacted] 16 days ago [redacted]

.github	[redacted]	7 months ago
.all-contributorsrc	[redacted]	5 months ago
.bumpversion.cfg	[redacted]	5 months ago
.editorconfig	[redacted]	9 months ago
.env.example	[redacted]	9 months ago
.gitignore	[redacted]	8 months ago

[redacted]

Figure 3.2: In red, one of the locations of the ".all-contributorsrc" file an OSS project stored on GitHub. Sensitive information has been masked.

[1] https://github.com/LABHR/octohatrack
[2] https://allcontributors.org/



dev-requirements.txt [redacted] 8 months ago
main.py [redacted] 5 months ago
renovate.json [redacted] 8 months ago
requirements.txt [redacted] 5 months ago
topics.txt [redacted] 9 months ago
utils.py [redacted] 16 days ago

README.md

[redacted] version [redacted] license [redacted] all contributors 7 Stars [redacted]

[redacted]

Usage

[redacted]

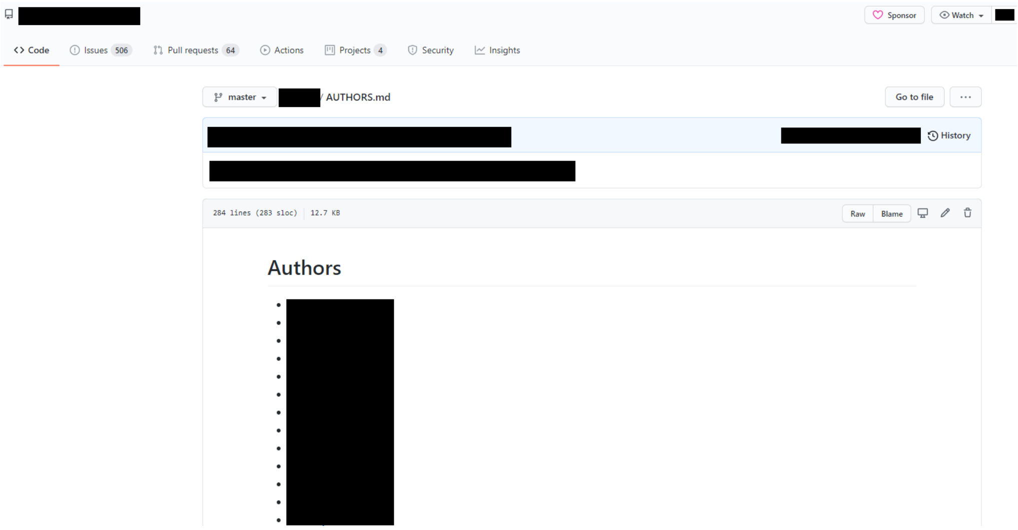
Contributors

Thanks goes to these wonderful people (emoji key):

[redacted]

This project follows the all-contributors specification. Contributions of any kind welcome!

Figure 3.3: In red, the content of the ".all-contributorsrc" file in an OSS project stored on GitHub. Sensitive information has been masked.



[redacted] / AUTHORS.md

[redacted] [redacted]

[redacted]

284 lines (283 sloc) | 12.7 KB

Authors

[redacted]

Figure 3.4: The content of the "authors" file in an OSS project stored on GitHub. Sensitive information has been masked.

We have decided, to focus on the last three models of contributorship to cover as many contributors as possible in an OSS projects. The first model was not taken into account in this research because of the GitHub limitation of showing only the most 100 contributors, by number of commits.

3.3 Data collection

OSS are hosted / stored on various online platforms, like GitHub[3], GitLab[4] and BitBucket[5]. Such platforms allow people to join and contribute in an OSS project. While this is a wonderful opportunity for people to get in contact and work together, it might also lead to offensive behaviour, as mentioned in Section 2.2. To determine, from a quantitative point of view, the impact that code of conduct has on women's involvement,

[3] https://github.com/
[4] https://about.gitlab.com/
[5] https://bitbucket.org/

it is important to identify as many contributors in OSS projects as possible. Thankfully, online platforms, such as GitHub, provide a lot of information with respect to who has contributed on a project, when and how. For this thesis, only the data from OSS projects stored on GitHub is considered. The reason for studying only OSS projects is that we do not have access to non-OSS projects stored on GitHub stored. The reason for choosing GitHub is that GitHub is the largest online platform for storing OSS projects, having more than 96 million of such projects[6].

Besides determining where to extract the data from, it is important to understand what data has to be extracted from GitHub repositories to answer **RQ1** and **RQ2**, defined in Section 1.1. Since the research questions are similar, the same following information is needed:

- List of OSS projects which have adopted a code of conduct;
- Information about the people who have contributed to the OSS projects which have adopted a code of conduct, based on which contributor's gender can be determined. The information needed to identify the gender of contributors is described in Section 3.3.1. In essence, the gender of the contributors is determined based on their name. For identifying the gender based on name, Namsor[7] is used. The reasons for choosing Namsor are also presented in Section 3.3.1;
- List with the dates when each of the people, previously identified, has contributed. Such list is created for each of the OSS projects which have adopted a code of conduct. The dates are used to determine if the contribution was done before or after the code of conduct was introduced in the OSS project, useful for comparing the contributions done in the pre-conduct period with the ones done in the post-conduct period. Furthermore, the dates are used to determine how many women have contributed in the OSS projects and for how long.

3.3.1 Data and tool for determining gender

For identifying the gender of a contributor in an OSS project, various methods have been attempted before:

- Identifying gender based on name: this method requires extracting the names of contributors, and based on the name, either identify the gender manually, or use a tool for gender identification, such as genderComputer [54];
- Identifying gender based on profile image: this method requires extracting the profile image of contributors, and based on the image, either identify the gender manually, or use a tool for gender identification, such as the one described by Lu et al. [28]. Problem with this approach is that GitHub profile pictures might not represent facial images, leading to wrong classification of gender. Furthermore, images might need manual preprocessing, to rotate or crop images before identifying the gender [54]. Another problem is that "GitHub profile pictures are scarce" [42]. To determine how scarce GitHub profile pictures are, we have picked a sample (Confidence Level 95% and Confidence Interval 5%) consisting of 384 users, from the 166,597 users presented in Table 3.8, and manually analyzed how many users have a name and how many users have a picture representing a human face. Out of 384 users, 331 users have a name (86.2% out of all users) while only 107 (27.86% out of all users) had a picture representing a human face. As a result of the manual analysis, we can indeed confirm that GitHub profile pictures are few in comparison to names. Because of the small percentage of GitHub profile pictures, we have decided to discard this approach;
- Identifying gender based on artefacts: this method requires extracting and analysing contributor's messages [2], and based on the content and writing style, a gender is determined.
  - Identifying gender based on the content: users can identify the gender of the people with whom they communicate online, based on the knowledge that people have on various topics, e.g., related to ring sizes and pantyhoses [50]. One problem with this approach is that it cannot be applied in a multicultural context, due to the differences in cultures [29]. Another problem is that the communication is topic-restricted in online software communities, and more personal topics, based on which a gender could be identified, are not allowed [54]. A topic-restricted communication

[6] https://octoverse.github.com/

[7] https://www.namsor.com/

might explain why a study by Kruger and Hermann [26], has revealed that the research for identifying gender based on text, is focused only on twitter posts, facebook posts, news articles and novels.

– Identifying gender based on writing style: The study by Argamon et al. [2], has shown that there are gender stylistic language features based on which gender can be determined, e.g., women use pronouns such as "I", "you" and "she" more often than men. The problem with this approach is the absence of gender stylistic language features, as shown by Herring and Paolillo [20]. The work done by Herring and Paolillo [20] has shown that the writing style has an impact of gender stylistic features, independent of the author gender, e.g., diary has more 'female' stylistic features, in comparison to a filter blog. The work done by Argamon et al. [2] is focused on research papers and on fiction documents, while the writing style on GitHub might influence the style-based gender-resolution.

Given the limitations of identifying gender based on profile image and artefacts, we have decided to identify gender based on the name. This approach has been used in other research papers which study the gender diversity in GitHub [55–57]. The name of the people, who have contributed to the OSS projects which have adopted a code of conduct, is extracted from GitHub.

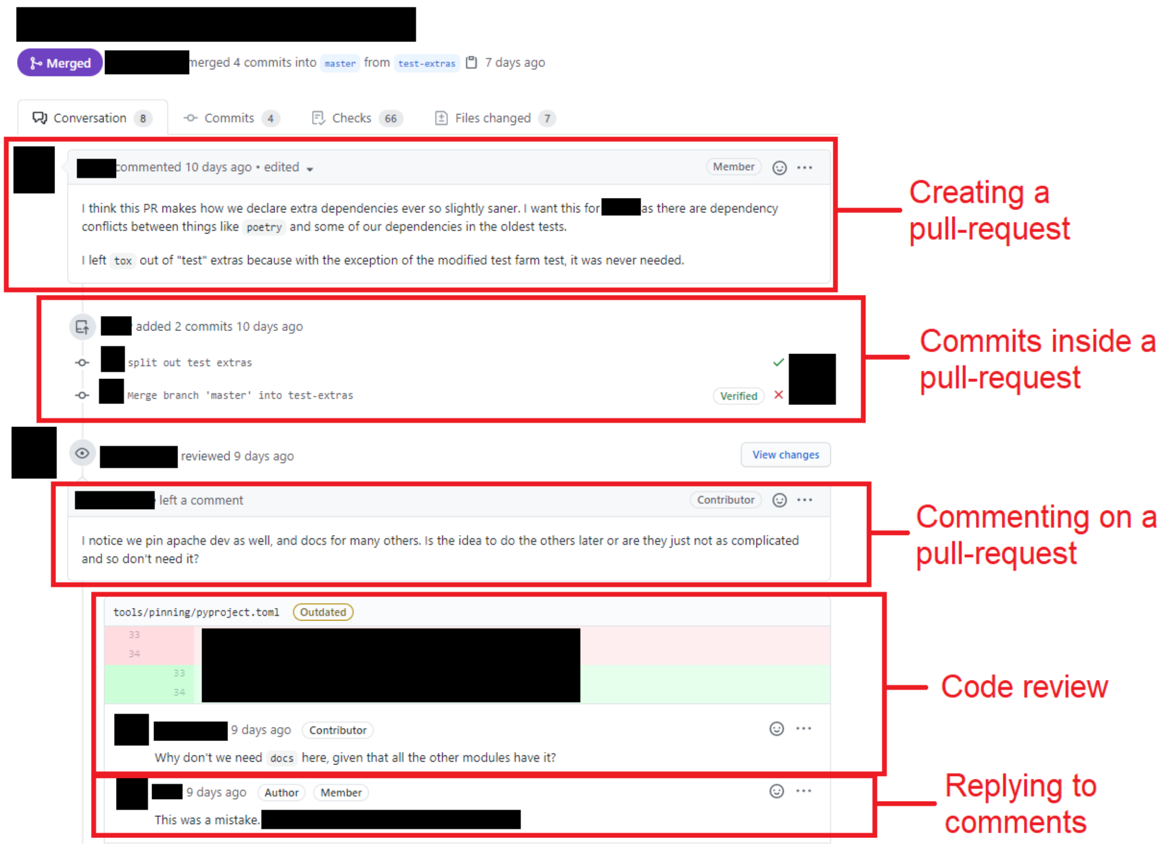
Because of the large number of contributors to identify gender, more than 140,000, as shown in Table 3.9, deciding manually the gender for each of the contributors would require too much time. Instead, a tool, for automatically identifying names is used. For this thesis, Namsor[8] is used for identifying gender based on name. Namsor is a powerful tool which uses machine learning for determining the gender based on names. The advantage of using Namsor, in comparison to other tools such as Gender-API[9], is that the name does not have to be split into first name and last name, because Namsor automatically identifies the first, middle (if applicable) and last name. The tool is not case sensitive, and also deals with abbreviations, such as *B. Enache* for example. Furthermore, it also supports Asian names, which are always tricky to infer gender from because of the different alphabets used. Namsor also takes into account the location of the name, if one is provided, which is useful since names can have different genders based on the location, e.g., "Andrea" which in Italy is a male name while in the rest of the world is a female name.

3.4 GitHub platform, contribution and features

GitHub is one of the largest online platforms, for OSS projects, in the world, allowing people and organizations to store OSS projects and work together on GitHub. Storing OSS projects on GitHub, allows people, who have the right access to the project, to modify any file of the project, keep track and revert changes if needed. To modify a file from a repository, a commit needs to be made, containing the file changes, and pushed to the repository. Depending on how the GitHub repository is set, the changes can be either applied directly to the repository, after a push is done, or the changes might require approval of other community members, in the form of a pull request. The pull-requests can contain, besides one or more commits, comments, replies to comments and comments on code, as shown in Figure 3.5.

[8] https://www.namsor.com/

[9] https://gender-api.com/



[redacted]

Merged [redacted] merged 4 commits into `master` from `test-extras` 7 days ago

[redacted] [redacted] commented 10 days ago • edited Member

I think this PR makes how we declare extra dependencies ever so slightly saner. I want this for [redacted] as there are dependency conflicts between things like `poetry` and some of our dependencies in the oldest tests.

I left `tox` out of "test" extras because with the exception of the modified test farm test, it was never needed.

Creating a pull-request

[redacted] added 2 commits 10 days ago

[redacted] `split out test extras` ✓ [redacted]

[redacted] `Merge branch 'master' into test-extras` Verified ✗

Commits inside a pull-request

[redacted] [redacted] reviewed 9 days ago View changes

[redacted] left a comment Contributor

I notice we pin apache dev as well, and docs for many others. Is the idea to do the others later or are they just not as complicated and so don't need it?

Commenting on a pull-request

`tools/pinning/pyproject.toml` Outdated

[redacted]

[redacted] [redacted] 9 days ago Contributor

Why don't we need `docs` here, given that all the other modules have it?

Code review

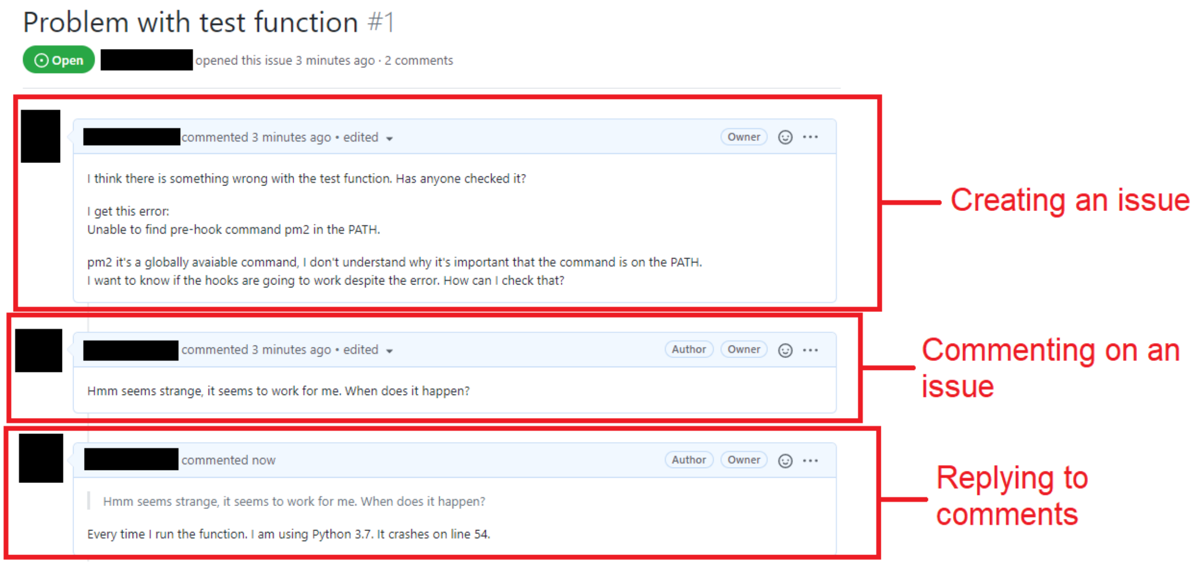
[redacted] [redacted] 9 days ago Author Member

This was a mistake. [redacted]

Replying to comments

Figure 3.5: The types of pull-request interactions available on GitHub

In addition to commits, pull-requests, pull-requests comments, pull-requests replies and pull-requests code comments, people can report errors, propose new features / ideas, or ask questions, by creating an "Issue" in the repository. People can comment and reply in an issue, as seen in Figure 3.6.



Problem with test function #1

Open [redacted] opened this issue 3 minutes ago · 2 comments

[redacted] [redacted] commented 3 minutes ago • edited Owner

I think there is something wrong with the test function. Has anyone checked it?

I get this error:
Unable to find pre-hook command pm2 in the PATH.

pm2 it's a globally avaiable command, I don't understand why it's important that the command is on the PATH.
I want to know if the hooks are going to work despite the error. How can I check that?

Creating an issue

[redacted] [redacted] commented 3 minutes ago • edited Author Owner

Hmm seems strange, it seems to work for me. When does it happen?

Commenting on an issue

[redacted] [redacted] commented now Author Owner

> Hmm seems strange, it seems to work for me. When does it happen?

Every time I run the function. I am using Python 3.7. It crashes on line 54.

Replying to comments

Figure 3.6: The types of issue interactions available on GitHub

Pull-requests, issues, commits, comments, replies to comments and comments on code are also known as GitHub interactions [62]. Each of the aforementioned interactions is stored on GitHub, in the GitHub repository, containing the date of the interaction, the actor who has made the interaction as well as the content of the interaction. The interaction data is accessible on GitHub, via the GitHub UI, or via the GitHub API. Based on the definition of contributorship, described in Section 3.2, all of the interactions just described above represent the model 2 of contributorship.

Since people can contribute to a project in ways different than the ones just presented, such as creating events, helping with translation, making posters, their contribution will not be shown in the GitHub repository. To solve this problem, the owners or maintainers of a GitHub repository can create, at their own initiative a list with all the people who have contributed in the repository. This list is stored in a file, usually with a name such as “authors” or “contributors”, inside the GitHub repository, as shown in Figure 3.4. GitHub does not have any standard format for giving credits to people who have contributed in other ways than the ones just mentioned. The problem with such files is that they do not have a standardised format, making it very difficult to extract the list of contributors from the file, because other information might be found in the file (such as a thank you message, or a table, or information about how to contribute). Some maintainers opt to use a standardised format created by the All-Contributors community, from which data can easily be extracted. As explained in Section 1.1, these files represent the models 3 and 4 of contributorship.

For extracting the data for this thesis, we have used the GitHub API[10] since it allows for large data to be extracted from GitHub. Another option for extracting data from GitHub is GHTorrent, a dataset, made publicly available, that is a mirror of GitHub [19]. GHTorrent is split into two: a MySQL dump, which contains all the meta-data for the GHTorrent dataset, and a MongoDB dump, which contains all the textual data. Currently, the MongoDB has 18TB worth of data, while the MySQL dump has 146GB. For this thesis, we have decided to not use GHTorrent because identifying projects which have adopted a code of conduct, requires both the MySQL dump (to determine when a contribution took place) and the MongoDB dump (to determine whether the contribution was related to adopting a code of conduct). Furthermore, identifying the users who have contributed in a project, requires both datasets, one for determining when the contribution took place, and the other one containing information about who has made the contribution. The amount of storage required for identifying projects which have adopted a code of conduct, as well as the contributors for these projects, is enormous. Furthermore, the latest available MongoDB dump is from June 2019[11], and we would like to extract data until the pandemic has started (March 1^{st}, 2020). Since there is no publicly available list with GitHub repositories which have adopted a code of conduct, the only way for identifying such repositories is by using the GitHub search function, provided in the GitHub API.

Instead of using GitHub API, PyDriller could be used as well. The reason for not choosing PyDriller in this thesis is related with the GitHub search limitation. When searching for repositories, such as GitHub repositories which have adopted a code of conduct, GitHub limits the results to only 1000 answers. PyDriller does not offer a solution for avoiding this limitation.

3.5 Repository selection

At the time of writing this thesis, there is no list of GitHub repositories which have adopted a code of conduct. GitHub offers their users the option to add a code of conduct when creating a repository[12], but it does not disclose repositories which have adopted the code of conduct.

The repositories which have adopted a code of conduct are the ones which have a code of conduct file. To identify the repositories which have such files, the search engine built inside GitHub is used. Similar to the work done by Tourani et al. [52], searching for repositories which have a code of conduct is done using the query *“code of conduct”*. By default, GitHub searches based on commits, thus, all the results returned by GitHub were commits which contained the words *“code of conduct”* in the body of the commit. Using this search strategy, results returned were inconsistent: the same query can return 5,000 results or 500,000 results. One possible reason for this inconsistency is the way GitHub updates and indexes the commits.

[10] https://docs.github.com/en/rest

[11] https://ghtorrent.org/downloads.html

[12] https://docs.github.com/en/communities/setting-up-your-project-for-healthy-contributions/adding-a-code-of-conduct-to-your-project

Because of this inconsistency, it was decided to search by filename, in order to find projects which have adopted a code of conduct file.

To identify the filenames for files which have a code of conduct, a search by code, using the query "code of conduct" was performed. This search result returns all files which contain the words "code of conduct" inside the file, or in the title of the file. A limitation of GitHub search is that, no matter the type of search used, GitHub will return only the first 1000 results. Because of this limitation, search was split based on the size of the files which contain the words "code of conduct" in order to circumvent the GitHub limitation. In the end, the following query was used:

"code of conduct" size:XX..YY

where XX and YY represent the file size (in bytes) of the files which contain the "code of conduct" words. When extracting the data, the XX and YY values are always incremented after extracting all the GitHub results from the current query. This process is repeated until no more results are found. For example, searching by code using the query *"code of conduct" size:106..107* will return all files containing the words "code of conduct", that have a file size of 106 or 107 bytes. Furthermore, the repository in which the file is stored is also returned. This "slicing" idea has been used in other papers, such as the one by Young et al. [62], in which the different types of contributions are analyzed.

All the results were extracted from GitHub and the results returned were ranked by the number of results each file name had. A random uniform sample, confidence level 95%, confidence interval 10%, was picked for the ten most common file names in the repositories found using the queries "code of conduct size:XX..YY" to determine how many of these file names are actually a code of conduct. In this way the accuracy for each of the most common ten file names could be determined. The rank with the most common file names can be found in Appendix A.1. The majority of the GitHub results are in files named "readme", "code of conduct", "contributing", "changelog" and "conduct". Another interesting fact is that all of them have the extension "md".

After manually analysing each of the samples for the most ten common file names, the conclusion was that no files named "readme", "contributing" and "changelog" are actually a code of conduct. Instead, the files named "readme", "contributing" and "changelog" are just mentioning the code of conduct, while all files named "code of conduct" and "conduct" had a code of conduct inside them. Based on these findings, the filenames used for identifying the code of conduct files inside a project are "code of conduct" and "conduct". In the end, the following queries were used to identify code of conduct files:

filename:"code of conduct" extension:md size:XX..YY
filename:"conduct" extension:md size:XX..YY

where XX and YY represent the file size (in bytes) of the code of conduct file to be searched for. When extracting the data, the XX and YY values are always incremented after extracting all the GitHub results. This process is repeated until no more results are found. For example, searching by code using the query *filename:"code of conduct" extension:md size:106..107* will return all files named *code of conduct*, that have a file size of 106 or 107 bytes and ending with the extension md. Furthermore, the repository in which the file is stored is also returned. It is important to mention that the GitHub search is not case sensitive and if the search is "code of conduct", files with names "CoDe-of_ConDUCT.mD" can still be found.

While identifying the projects which have a code of conduct, it was noticed that GitHub returned multiple results for the same project. After manually looking at the duplicates, it was concluded that a project can have multiple code of conduct files. The reason for this is that a projects have external libraries added in the repository, such as ruby[13] and phpunit[14], which also have a code of conduct. To solve the problem, a threshold for the path depth needs to be set such that every code of conduct file with a path longer than the threshold is considered to be part of an extension representing other projects. Thus, if the code of conduct file has a path longer than the threshold, it is considered that the code of conduct does not belong to the current project. If the current project does not have any other code of conduct file, it is simply removed from the dataset.

[13] https://www.ruby-lang.org/en/
[14] https://phpunit.de/

The reason for choosing a threshold instead of removing duplicate repositories in the post-processing is that there are projects which do not have a code of conduct file, but which use an external library which has a code of conduct. Such repositories are wrongly identified as having a code of conduct, when in fact, they do not have. Using the threshold approach, duplicates are removed, as well as repositories which are wrongly identified as having a code of conduct.

To compute the threshold, the path depth for each of the identified code of conduct files are computed and samples, confidence level 95%, confidence interval 10%, are taken from each of the path levels in order to determine how many are part of other projects. A list with how many results GitHub found for each of the path depths is available in Appendix A.2. The samples are manually analysed and the conclusion is that all code of conduct files with path depth longer than one, are part of external libraries. Thus, the threshold was set at one. Any code of conduct with path depth longer than one, is excluded from this analysis. For code of conduct files with a path depth of one, there are still code of conduct files that belong to external libraries, thus a list of good paths has to be created. To create a list of good paths, a sample size (confidence level 95%, confidence interval 10%) is taken for each of the 30 most common paths of length zero and one. Each of the sample sizes is analyzed to determine its accuracy (i.e. how many code of conduct files that are in path X are actually belonging to the project itself). Such list is available in Appendix A.3. The path depth of zero are also analysed just to make sure that code of conduct files stored in the root file (i.e. path depth zero) are belonging to the project itself and not to external ones. Based on the 30 most common path files for code of conduct, it was concluded that the root path (i.e. path depth zero) has accuracy of 100%. For paths with depth one, the paths ".github" and "docs" have accuracy of 100% while the other ones have accuracy of 0%. The root, ".github" and "docs" paths are exactly the same as the ones recommended by GitHub to store a code of conduct in a project[15].

In conclusion, finding repositories which have adopted code of conduct is done using the following GitHub by code queries:

filename:"code of conduct" extension:md size:XX..YY path:/
filename:"code of conduct" extension:md size:XX..YY path:/.github
filename:"code of conduct" extension:md size:XX..YY path:/docs
filename:"conduct" extension:md size:XX..YY path:/
filename:"conduct" extension:md size:XX..YY path:/.github
filename:"conduct" extension:md size:XX..YY path:/docs

The statistics for the process of identifying, on GitHub, the repositories which have a code of conduct, are found in Table 3.1.

Results GitHub found	139,486
Results retrieved	121,972
Repositories retrieved	120,665
Results missed because of GitHub limitations	17,514

Table 3.1: Overall statistics obtained after identifying the repositories which have a code of conduct using the aforementioned queries

Unfortunately, retrieving the missed GitHub repositories because of GitHub limitations is not possible. GitHub does not offer any paid service to raise the search limitation of 1000 results. Another option to circumvent the 1000 results search limitation would have been to split the results based on date a file was modified. For example, search for all code of conduct files which were modified between date X and date Y. Unfortunately, at the time of writing this thesis, searching for file modifications in a certain interval does not work, GitHub searching tool returning 0 results. GitHub also provides an option to search by number of stars. Unfortunately, searching by stars is restricted only when searching by repositories. Searching by repositories does not support queries of the form *filename:"code of conduct"*, meaning that we cannot find

[15] https://docs.github.com/en/communities/setting-up-your-project-for-healthy-contributions/adding-a-code-of-conduct-to-your-project

repositories which have a code of conduct, we can only find repositories which have between X and Y stars. These are the only advanced searches that GitHub supports at the time of writing this thesis. Due to the lack of any other search options, which might circumvent the search limitation of 1000 results, there is no way to refine the queries and avoid missing the results due to GitHub limitations.

3.6 Repository filtering

The dataset obtained in Section 3.5 is filtered to eliminate external factors as much as possible. It is important to mention that each filtering step is applied on the dataset obtained from the previous filtering process.

3.6.1 Filtering Covid threats

Since this thesis is done during a pandemic, the threat of Covid influencing our results was taken into account, especially since Covid has had (and still has) a massive impact on everyone's life. Covid has greatly changed the way people work [30, 35, 38, 47]. Software developers have also changed their ways of working because of the pandemic [60].

Taking into account the disruption that Covid had on everyone's life, it was decided to not extract any data that was generated after 01-March-2020, the date when the pandemic was officially declared by World Health Organization[16]. This date represents the end-point of this research. A consequence of this decision is that all repositories which have adopted a code of conduct after 01-March-2020 have to be removed from the dataset.

The statistics for this filtering step are found in Table 3.2.

Number of repositories removed	45,221
Number of repositories kept	75,444

Table 3.2: Overall statistics for removing repositories which have adopted a code of conduct after the pandemic started

3.6.2 Filtering based on contributorship models

As discussed in Section 1.1, identifying the contributors for each repository is done based on the model 2, 3 and 4 of contributorship defined by Young et al. [62]. Using these models requires each repository to fulfill model 2 and either model 3 or 4 of contributorship in order to ensure that no contributors are missed. As described in Section 3.2, model 2 of contributorship is related with the contributors identified based on the interactions with the repository (commits, pull requests, issues, discussions, comments, replies, reactions and so on), model 3 of contributorship is related with contributors identified based on the ".all-contributorsrc" files, and contributors that can be identified by parsing non-standardized data sources (i.e., "contributors" and "authors" files). All the remaining repositories from filtering step described in Section 3.6.1 fulfill the model 2, since all of them are GitHub repositories and all of them have at least one GitHub interaction (otherwise they would have not been found in the GitHub search). For fulfilling model 3 or 4, repositories need to have either a standardized file (as required by model 3) which contains a list of contributors, as shown in Figures 3.2 and 3.3, or a non-standardized file (as required by model 4) which contains a list of contributors, as shown in Figure 3.4. For model 3, the only standardized file at the time of conducting this research, is the ".all-contributorsrc" file created by the All-Contributors community. For model 4, the challenge is in identifying what file names are used for the list of contributors. For this thesis, the "authors" and "contributors" file names are used for model 4. Both of these names were used in the paper by Young et al. [62] as well as by the octohatrack tool [17], a popular software used for identifying all the contributors from a GitHub repository. Thus, the repositories remained after the filtering step described in Section 3.6.1 are checked for the existence of an ".all-contributorsrc", "authors" or "contributors" file. The repositories which do not have such files are removed.

[16] https://www.who.int/

[17] https://github.com/LABHR/octohatrack

The statistics for this filtering step are found in Table 3.3.

Number of repositories with ".all-contributorsrc" file	888
Number of repositories with "authors" file	3902
Number of repositories with "contributors" file	1134

Table 3.3: Overall statistics with repositories that have an ".all-contributorsrc", "authors" or "contributors" file

Furthermore, the files which are not text were removed. This step was done manually because of the low number of repositories (as shown in Table 3.3). The results from this filtering step are found in Table 3.4.

Number of repositories with ".all-contributorsrc" file	888
Number of repositories with "authors" file	3817
Number of repositories with "contributors" file	1077

Table 3.4: Overall statistics with repositories that have an ".all-contributorsrc", "authors" or "contributors" file after removing files which are not text or which belong to external projects

It is important to mention that the projects from Tables 3.3 and 3.4 do not represent disjoint sets. Looking at the dataset it was concluded that there are repositories which have a "contributors" file as well as an "authors" file. The projects which have an "authors" and a "contributors" file were manually analysed and it was concluded that they either contain the same content, or that one file is a superset of the other, in both cases the files had the same update history. Either the superset was kept, in case there exist one, or one of the files were kept, in case they had the same content and history. The number of projects remained after this filtering step is 5556.

Like every other file, the ".all-contributorsrc", "authors" and "contributors" files have an update history. To determine if there is a difference, with respect to women's involvement identified in model 3 and 4 of contributorship, between the pre-conduct and post-conduct period, these files need to have at least an update before the pre-conduct and an update after the post-conduct. Otherwise, there is nothing to compare in model 3 and 4 of contributorship. Thus, another filtering step for model 3 and 4 is removing the repositories which do not have at least one model 3 or 4 update in the pre-conduct and at least one model 3 or 4 update in the post-conduct period. The results from this filtering step are found in Table 3.5.

Repositories before removing the ones which do not have at least an update for the model 3 or 4 of contributors in the pre-conduct and post-conduct periods	5556
Repositories after removing the ones which do not have at least an update for the model 3 or 4 of contributors in the pre-conduct and post-conduct periods	1715

Table 3.5: Overall statistics for removing repositories which do not have at least an update for the model 3 or 4 of contributors in the pre-conduct and post-conduct periods

The number of repositories which have a "contributors" file is 306, the number of repositories which have an "authors" file is 1264, and the number of repositories which have an ".all-contributorsrc" is 145.

3.6.3 Filtering cloned repositories

Having cloned repositories in the dataset might influence the results of this research. For identifying and removing clones, also known in GitHub terms as forked repositories, the two definitions given in the paper by Pietri et al. [40] are used:

- **Type 1 forks:** also known as forged forks, are made by clicking the "fork" button in GitHub

- **Type 2 forks:** a repository, B, is a type 2 fork, also known as a shared commit fork, of repository A, if there is a commit C which both repositories have. To identify a shared commit, the commit SHA [18] is used.

The paper also mentions that one cannot rely on identifying clones solely based on type 1 and type 2 forks. Thus, in case repository B is a type 2 fork of repository A, someone must manually look at them and determine if they are indeed forks or not, as well as determine which repository is the forked one and which is the original one. Thus, all the clones identified using the aforementioned definition are manually checked to determine if they indeed are clones, and establish which repository is the original one and which should be removed. Since a project might be cloned after a code of conduct was adopted, or before, it is important to split the clones in different cases when analysing them:

- **Case 1:** If project was cloned after the code of conduct was adopted: remove all its clones since all of them share the same pre-conduct period as the original project, leading to biases in the analysis;
- **Case 2:** If clone happened before the code of conduct:
  - **Case 2-1:** if project is a clone having an identical contribution history as the original one: remove the clone since it does not bring any new information and leads to biases in the analysis;
  - **Case 2-2:** if project is a clone having an identical contribution history as the original one right before cloning, but after cloning they have a totally different contribution history. The only two possible reasons why this happens are:
    - **Case 2-2-1:** the team has split: in this case, both project should be removed since it is impossible to keep track on which people remained in the original project and which moved to the spin-off
    - **Case 2-2-2:** someone cloned the project and started to contribute on the clone, independently of the initial team (i.e. the team did not split): for this case, both the original project and the clone are kept, but the measurements are done differently. For the original project, the entire history is kept, but for the cloned one, the contribution history is measured only after the last commit that was made before the project was cloned, since starting with the last commit, both projects have a totally different history.

To identify whether the repository was cloned before or after the code of conduct was adopted, the commit history of the code of conduct file from both the original project and its clones are manually analysed to determine if they have any common commit SHA. If they have any common commit SHA, then the project was cloned after the code of conduct was adopted. In case they do not have any common commit, the content of the code of conduct files is analysed to determine if they have the same contact information or project description. In case they have the same information, it means that the project was cloned after the code of conduct was adopted.

The number of clones found can be seen in Table 3.6.

Number of repositories before removing clones	1715
Number of clones in case 1	1176
Number of clones in case 2	0
Number of repositories remained after removing clones	539

Table 3.6: Overall statistics for removing clones

As can be seen in Table 3.6, more than half of the repositories from the initial dataset are clones. After manually looking at the dataset to understand why there are so many clones, it was concluded that the dataset has some popular GitHub repositories, such as TensorFlow[19], a very popular machine learning library, which attract a lot of interest from the OSS community, many people wanting to clone it in order to

[18] https://blog.thoughtram.io/git/2014/11/18/the-anatomy-of-a-git-commit.html

[19] https://github.com/tensorflow/tensorflow

either contribute or change / tweak some aspects from such popular repositories. For example, TensorFlow has more than 200 clones in the dataset.

3.6.4 Filtering non-engineered projects

Recent studies have shown that there are GitHub repositories which are not active nor collaborative. Such type of projects are mainly used as storage or for educational purposes [24] and have to be removed from the dataset since they do not contain frequent contribution which is needed to determine how women's involvement is influenced by code of conduct. A solution, for removing inactive project, found in many recent research papers, relies on applying a set of filtering criteria [15,21,37,46,58].

Besides applying a set of filtering criteria, there is another solution, in the form of an automated tool, presented by Munaiah et al. [31], which classifies each project in either a non-engineered project, or an engineered project. According to Munaiah et al. [31], engineered projects are collaborative and active software projects. Classifying the projects in the dataset is done by measuring various metrics for each of the GitHub projects, metrics such as the number of commits, their frequency, the presence of software tests and the number of contributors. All these metrics measure activity and community engagement. Based on the metrics computed, a classifier is used to establish whether a project is a engineered or non-engineered one. The main drawback of this approach is that the tool created by Munaiah et al. [31], named RepoReaper, is very slow, requiring days for analysing several hundred projects.

For this thesis, given the small number of projects in the dataset, as shown in Table 3.6, the RepoReaper tool was used. The number of engineered and non-engineered software projects can be seen in Table 3.7.

Number of engineered software repositories	418
Number of non-engineered software repositories	121

Table 3.7: Overall statistics for removing non-engineered software repositories

3.7 Identifying contributors

As presented in Section 3.2, a contributor is a person who does a contribution in an OSS project. To identify contributors, together with their contributions, the models of contributorship, described by Young et al. [62], are used. As described in Section 3.2, model 2 of contributorship is related with the contributors identified based on the interactions with the repository (commits, pull requests, issues, discussions, comments, replies, reactions and so on), model 3 of contributorship is related with contributors identified based on the ".all-contributorsrc" files, and contributors that can be identified by parsing non-standardized data sources (i.e., "contributors" and "authors" files). Since the analysis is based on comparing women's involvement from pre-conduct period, with the one from post-conduct period, the following information is needed from each contribution:

- The name of the contributor. Based on the name, the gender can be identified. The reason for relying on name to identify gender is that GitHub does not store such information on their platform;
- The location of the contributor. Since names can have different genders based on the location, for example "Andrea" which in Italy is a male name while in the rest of the world is a female name, it is best to also include the location for each of the logins which have a name;
- The date of the contribution. Based on the date, it can be determined if the contribution took place before or after the code of conduct was adopted in the project. Date also helps in determining how much women have contributed and for how long.

Since model 3 and model 4 of contributorship are just files which contain a list with the names of the contributors, but not the date / period when they have contributed, the date of the contribution for these people is the date when they were added in the list found in the ".all-contributorsrc", "authors" files. Because model 3 and 4 of contributorship do not have such a rich data as model 2 of contributorship, a different statistical model is used for them, as described in Section 3.10.

3.7.1 Identifying contributors based on model 2 of contributorship

Based on the definition of model 2 of contributorship, presented by Young et al. [62] the following GitHub interactions were extracted from each GitHub repository, using the GitHub API and GitHub GraphQL:

- the pull-request interactions, shown in Figure 3.5:
  - creating a pull-request
  - creating the commits included in the pull-request
  - commenting on a pull request
  - replying to comments made in a pull request
  - code review
- issue interactions, shown in Figure 3.6:
  - creating an issue
  - commenting on an issue
  - replying on comments made in an issue
- commits

The only interaction that is missing is reacting to a comment. The reason why it was not included is that the date of a reaction is not stored on GitHub and GitHub does not provide any API for extracting reactions. Besides this, all the interactions are complete.

An overview of all GitHub interactions extracted can be seen in Table 3.8.

Type of GitHub interaction	Number of interactions	Number of contributors
Pull-request	6,192,034	58,781
Issue	1,434,198	129,725
Commit	3,223,002	40,946
Total	10,849,234	166,597

Table 3.8: Overall statistics for model 2 of contributorship

As specified in Section 3.6.1, it is important to mention that because of the great impact that Covid had on the entire world, including software developers [60], no GitHub interactions made after the 01-March-2020 are extracted.

For each of the 166,597 contributors found based on the GitHub interactions, the name was extracted from their GitHub profile page. Since having a name on your GitHub profile is something optional, not all contributors identified have a name. Fortunately, GitHub stores the name of the contributors in the GitHub interactions as well, thus, the names were extracted from the interactions as well. In case the contributor still does not have a name, their interaction history is used, since they might contribute in projects that are outside the scope of this paper (i.e. projects which either do not have a code of conduct, or which were filtered out). In case a contributor has names both in the GitHub profile page and in the GitHub interactions, all the unique names are kept and assigned to the contributor as being their full name.

For preprocessing the names, a strategy similar to the one done by Vasilescu et al. [54], was applied to the names found. Special characters were removed from words[20]. Symbols were also removed from words[21]. Words were converted from Leet to Latin (e.g., "w4lt3r" stands for "walter") [44]. A list with blacklisted words was created. This list was created by using more than 16,000 email provider domains[22] and country

[20] https://owasp.org/www-community/password-special-characters

[21] https://www.alt-codes.net/

[22] https://gist.github.com/drakodev/e85c1fd6d9ac8634786d6139e0066fa0

domains[23]. Github assigned names, such as “github124”, were also added to the blacklist. While this list might not be complete, it should be enough to keep words which might represent potential words. If the full name is A B C, and B is blacklisted, the full name will be A C. If full name is B and B is blacklisted, that contributor (together with all of its contributions) is removed from the data. At the end, the contributors that have a bot name were removed. Identifying and removing bots was done using BoDeGHa and BIMAN tools [13, 18]. A manual check was done afterwards of the identified bots. The risk of missing bots is quite small, both models having an accuracy of 0.98% in identifying bots.

Since gender is inferred based on name, to increase the accuracy of the tool which determines the gender, the country location of the contributors is also determined in the same way as for names: based on the information available on the contributor’s GitHub user page. For the contributors which have a location on the GitHub user page, the country location was determined using Nominatim[24], a free geosearch tool.

The number of contributors with a name, for which a location was identified, can be seen in Table 3.9.

Number of contributors with name and who are not bots	142,602
Number of contributors with name, who are not bots and who have a location	85,299
Number of contributors who are bots	102

Table 3.9: Overall statistics for contributors of model 2 of contributorship

3.7.2 Identifying contributors based on model 3 and 4 of contributorship

Identifying contributors from model 3 and 4 of contributorship requires a different strategy than the one explained in Section 3.7.1. It requires extracting the names of the contributors, together with their location where applicable, from list stored in the “.all-contributorsrc”, “authors” and “contributors” files. A copy of the content of the files, after each update, was extracted from GitHub. With the files stored locally, the extraction of names from lists was done manually for model 4, because of the non-standard format of the files, as well as automatically for model 3, since, for the latter one, the format of the files is standard and the files are stored as JSON. It is important to mention that the contribution date assigned to each contributor name represents the date when the file was updated and name introduced. Because each update of a file contains the previous list together with the new names added to the list, the same contributor appears in multiple file updates. Because not all projects have women’s involvement in the pre-conduct or post-conduct period, as explained in Section 3.10.1, only 149 repositories, out of the 418 repositories obtained in Section 3.6.4, are used to answer RQ1. Model 3 and 4 of contributorship is not used to answer RQ2, since the list of contributors does not have information about the time when a contributor has left the project, as explained in Section 3.9.2. Since the process of identifying contributors is time consuming, mostly involving manual extraction, it was decided to identify the contributors from model 3 and 4 of contributorship only for the repositories that are used to answer RQ1.

As specified in Section 3.6.1, it is important to mention that because of the great impact that Covid had in the entire world, including on software developers [60], no “.all-contributorsrc”, “authors” and “contributors” file updates made after the 01-March-2020 are extracted.

The overall statistics with how many contributors, identified based on model 3 and 4 of contributorship, can be seen in 3.10.

Contributors with name and who are not bots	15,872

Table 3.10: Overall statistics for contributors of model 3 and 4 of contributorship

[23] https://en.wikipedia.org/wiki/Country_code_top-level_domain
[24] https://nominatim.openstreetmap.org/ui/search.html

3.8 Identifying gender

At this point, the dataset contains a list of contributors, including their full name, location (where applicable) as well as the dates when they contributed. As mentioned previously, GitHub does not contain any information about the gender of the contributors, thus the gender needs to be inferred based on the name and location. For this, Namsor[25] is used. For each contributor, a gender is determined with a certain degree of confidence. After a meeting with the creator of Namsor, it was advised to use a confidence level of 60% since this is a low enough threshold to infer gender for many contributors, but high enough to have a good accuracy (above 85%, as reported by the creator of Namsor).

The contributors for which a name could not be determined, or for which the confidence level was below 60%, a gender was not determined. A gender was determined for each of the contributors, identified in Section 3.7.1 and 3.7.2, who have a name.

An overview with the number of men and women contributors, together with their number of contributions, for model 2 are shown in Table 3.11 and for model 3 and 4 are shown in Table 3.12.

Number of repositories	418
Number of contributors	166,597
Number of contributors with no gender	25,560
Number of women contributors	10,230
Number of men contributors	106,812
Number of contributions	10,849,234
Number of contributions belonging to contributors with no gender	1,730,117
Number of contributions belonging to women contributors	921,199
Number of contributions belonging to men contributors	7,824,463

Table 3.11: Overall gender statistics for contributors of model 2 of contributorship

Number of repositories	135
Number of contributors	15,872
Number of contributors with no gender	1,476
Number of women contributors	753
Number of men contributors	13,643

Table 3.12: Overall gender statistics for contributors of model 3 and 4 of contributorship

3.9 Formulating hypotheses for RQ1 and RQ2

In this Section the null hypotheses for answering **RQ1** and **RQ2** are formulated, while taking into account the different contributorship models defined by Young et al. [62].

3.9.1 Formulating hypothesis for RQ1

The **RQ1** for this thesis is *Do projects have a (higher) increase in the proportion of women who contribute in the project, in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?*. Since this research is quantitative, a hypothesis needs to be formulated in order to be tested using a statistical model. Since the focus of this research is on determining whether the code of conduct has an impact or not on the proportion of women, the null hypothesis should assume that code of conduct does not have any impact on the proportion of women, and be rejected, or not, by a statistical model. The null hypothesis, **H0**, for **RQ1** is:

Adopting a code of conduct does not have any influence on the proportion of women who contribute in a project

[25] https://www.namsor.com/

To test this hypothesis, as mentioned in Section 3.1, the dataset needs to be processed and split into two: people who have contributed in the projects before the code of conduct was adopted, and people who have contributed in the projects after the code of conduct was adopted. In this way, there will be two categories of people who contributed, one which does not have any code of conduct “influence”, since it was not present in the project at the time they contributed, and one which is under the “influence” of code of conduct. Comparing these two categories, with respect to proportion of women, might determine whether the code of conduct has an influence on proportion of women who contribute. The split point in the dataset is when the code of conduct was adopted for the first time in the project. The split point is different to each project, since not all projects have adopted the code of conduct at the same time. Because of this, the time lengths of the pre-conduct and post-conduct periods might differ for each project. To ensure that the data is not skewed, it is important to have the pre-conduct and post-conduct of equal lengths for each project and across projects.

Throughout this chapter, whenever “time length period” is mentioned, it refers to both the time length of the pre-conduct period and the time-length of the post-conduct period, since they both need to have the same time length.

In case the null hypothesis is rejected by the statistical model, it can be concluded that the code of conduct has an influence on the proportion of women who contribute in the project. Looking at the coefficient for code of conduct, returned by the statistical model, it can be determined whether the impact is positive or not. If it is positive, it means that the proportion of women increases after the code of conduct is adopted, or decreases, if the coefficient is negative. In this way, an answer is given to whether the proportion of women increases in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted.

Making the pre-conduct and post-conduct period of equal time lengths for each project requires discarding a part of the contribution from the period that is longer. The time length period for a project is the time length of the shorter period. For example, if project A is four years old and has adopted the code of conduct three years after it was created, this implies that the pre-conduct period is three years and the post-conduct period is one year. The time length for each of the periods should be one year, since this is the time length of the shorter period (here post-conduct). To make the comparison fair between the two periods project A and B only one year of contribution from the pre-conduct period is taken into account from the pre-conduct period, representing the last year before the code of conduct was adopted. This time period is adjusted for each project in the dataset. The reason for using the last year is because the contribution is newer and has continuity with the post-conduct period, avoiding thus gaps in the analysis. For the post-conduct period, the entire year of contribution from post-conduct period is taken into account. Having equal pre-conduct and post-conduct periods and measuring only the contributions that are inside these two periods, is done for every project.

Making the pre-conduct and post-conduct period of equal time lengths across all projects is not straight forward. Taking the time length of the shortest period across all projects might lead to insufficient data to measure. For example, if project A has the time length period one month, project B has the time length period of three years and project C has the time length period of two years, it would not make sense to pick one month as a time length period, since it would imply to analyse only one month before the code of conduct was adopted and one month after the code of conduct was adopted. Since changes such as a code of conduct might take time to come in full effect [48], it might require a change in the behaviour of some communities behaviour, analysing only one month in each period does not reveal the impact that code of conduct might have in the project. Instead, a better solution is to keep the time length of two years and remove project A, since keeping it would lead to biases, because there is simply not enough data for project A. Deciding on a time length across all projects requires the visualisation of data. For this, more information is described in Section 3.10.1.

To remove any unclarities, each term used in the **RQ1** and in **H0** is specified below:

- **Projects:** open source projects hosted on GitHub. For this RQ1, the projects obtained in Section 3.8 are used;

- **Adopt code of conduct:** a code of conduct or conduct file was added to the GitHub project in paths “root”, “/docs” and “/.github”;
- **Pre-conduct period:** starts X days before the project adopted a code of conduct and ends one day before the code of conduct was adopted. The ends of the periods are also included in the period. X is defined in Section 3.10.1 since it needs to be the same across all projects;
- **Post-conduct period:** starts the day in which the project adopted a code of conduct and ends X days after the code of conduct was adopted. The start day is included in the period but the end day is excluded, in order to ensure that the post-conduct period has the same length as the pre-conduct period. X is defined in Section 3.10.1 since it needs to be the same across all projects;
- **Contribution and contributors:** as mentioned in Section 1.1, the model 2, 3 and 4 of contributorship are used to identify contributors and when they have contributed;
- **An increase:** higher / positive slope increase in the post-conduct period, in comparison to the pre-conduct period;
- **Proportion of women:** the number of women who contributed over a certain period in a project divided by the number of users who contributed over the same period in the same project.

3.9.2 Formulating hypothesis for RQ2

The **RQ2** for this thesis is *Do women remain contributors over a longer duration of time in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?.* Since this research is quantitative, a hypothesis need to be formulated in order to be tested using a statistical model. Since the focus of this research is on determining whether the code of conduct has an impact or not on the time women remain contributors in a project, the null hypothesis should assume that code of conduct does not have any impact on the time women remain contributors in a project, and be rejected, or not, by a statistical model. The null hypothesis, **H0**, for **RQ2** is:

Adopting a code of conduct does not have any influence on the time women remain contributors in a project

To test this hypothesis, as mentioned in Section 1.1, the dataset needs to be processed and split into two: people who have contributed in the projects before the code of conduct was adopted, and people who have contributed in the projects after the code of conduct was adopted. In this way, there will be two categories of people who contributed, one which does not have the “influence” of code of conduct, since it was not present in the project at the time they contributed, and one which is under the “influence” of code of conduct. Comparing these two categories, with respect to time women remain contributors in a project, will tell whether the code of conduct has an influence on how much time women remain contributors in a project. The split point in the dataset is when the code of conduct was adopted for the first time in the project. The split point is different to each project, since not all projects have adopted the code of conduct at the same time. Because of this, the time lengths of the pre-conduct and post-conduct periods might differ for each project. Since the time lengths of the pre-conduct and post-conduct periods are different, the maximum time a contributor can contribute in any of the two periods, is different.

The time length differences between the pre-conduct and post-conduct period might influence the results. For example, let A be a project which was created six years ago and which adopted a code of conduct two years ago. The pre-conduct period for project A is four years, while the post-conduct period is two years. Let X be a contributor who has contributed only in the pre-conduct period, and let Y be a contributor who has contributed only in the post-conduct period. When looking at the time these contributors have remained in the project, X was a contributor for two years and Y was a contributor for one and a half years. If the analysis is based on time length, X has contributed over a longer duration of time, in comparison to Y. The problem with this approach is that it does not take into account the time lengths of the pre-conduct and post-conduct periods. To take into account the time lengths of the pre-conduct and post-conduct periods, the time duration a contributor remains in a project is divided by the time length of the pre-conduct or post-conduct period, depending in which period they has contributed. Thus, instead of using time length, as a measure of unit, percentages, out of time length for pre or post-conduct period, are used. In the previous example, X has contributed 50% out of the pre-conduct period length, while Y has contributed 75% out of the post-conduct period length. Y has stayed longer in the project in comparison with X. Thus, when doing statistical analysis to determine how long contributors have stayed in a project, instead of using time length, percentages are used.

Another aspect that might influence the results is that not all contributors join a project at the same time. The problem is that people who have joined the project at a later stage, in any of the two periods, are at a disadvantage in comparison to people who have joined the project earlier. Taking the previous example, let's say that X has joined the project one year after it was created, and Y has joined the project right when the code of conduct was adopted. Despite the time length for the pre-conduct period being four years, X was not part of the project in the first year of the pre-conduct period, thus it would be unfair for them to take into account the entire four years when determining how long they has contributed. To fix this problem, the time length period is adjusted for each contributor, for each project. For X, since they started one year after project A was created, the time length for the pre-conduct period is three years, since the first year is excluded. For Y, since they have started right when the code of conduct was adopted, the length of the post-conduct period remains two years. If Y would have joined the project one year after the code of conduct was adopted, the post-conduct period would be one year for them.

In order to compute the time a contributor remained in a project, the date when they joined a project and left the project are needed. Unfortunately, for contributors identified in model 3 and 4 of contributorship, computing the time they remained in a project is not possible since the data from these models does not contain any information about the date when contributors left the project, or when they have contributed. Thus, for answering **RQ2**, only model 2 of contributorship is used. The date when a contributor joined a project is the date of the first model 2 contribution made by them in the respective project. Determining the date when a person has left the project is done by identifying the last contribution made by them in the respective projects. It was decided, in consultation with the supervisors, that the date when the person has left the project to be extended by a certain number of days, since the person might still be active in the project despite making no model 2 contributions. These added days are computed for each person in each project, by taking the longest time length between two consecutive contributions the respective person made in the respective project. After the dates when a person joined and left the project are determined, the number of days between these two dates are computed, and then transformed in percentages as explained in the last example.

People who, in a project, contribute in both periods are excluded. The reason for excluding them is because these people were influenced by the absence of a code of conduct as well as the presence of one. People who contribute beyond 01-March-2021 are excluded as well, since this is the time when the pandemic started. Only women and their contributions, are kept in the dataset obtained in Section 3.8.

In case the null hypothesis is rejected by the statistical model, it can be concluded that the code of conduct has an influence on the time women remain contributors in a project. In this way, an answer is given to whether women remain contributors over a longer duration of time in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted.

3.10 Statistical techniques for RQ1

To test the null hypothesis mentioned in Section 3.9.1, a suitable statistic technique needs to be selected. Based on the statistical technique, a statistical model is constructed to test the null hypothesis. Before selecting anything, it is important to understand what does adoption of a code of conduct implies. Adopting a code of conduct might change the behaviour of a community in an OSS project. This change might not happen overnight, it might be that the community needs time to embrace it, some people might leave, some people might join, depending on how they feel about the adoption of the code of conduct. It might take time to change some communities and make it more friendly towards minorities. Thus, the technique used should take into account the passing of time.

A technique which takes into account the passing of time is Regression Discontinuity Design (RDD) [9,51]. Adapting it to the purpose of this research would imply to determine how the proportion of women evolved over time in the pre-conduct period, and how it evolved in the post-conduct period (i.e. after the code of conduct was adopted in the project). In essence, two linear regressions are applied on the dataset: a linear regression in the pre-conduct period and a linear regression in the post-conduct period. The slopes of the two linear regressions are compared, together with the intercept. Any difference seen might be due to the code of conduct. Since RDD is focused on how the proportion of women evolves over time, repeated measurements of this metric need to be taken, over fixed interval. This fixed interval is also known as a time point [22]. For each time point, the proportion of women who have contributed in that time-point is computed. The time length of each time point should be equal in order to remove biases, since each time point is equally as important for determining the impact that code of conduct has on the proportion of women. In order to apply RDD, the time length period needs to be large in order to have enough time points to apply a linear regression. The number of time points for each period is computed by dividing the time length of the period to the time length of the time point.

For example, if the time length period is one year and the length of the time point is one month, it would mean that the contribution, from one year before the code of conduct was adopted and one year after the code of conduct was adopted, is analysed. Since the length of the time point is one month, it would mean that the proportion of women is computed for every month out of the 24 months in total. Thus, for every month, the proportion of women is obtained by counting how many women have contributed in that month and dividing it by the number of users who have contributed in that month. The pre-conduct and post-conduct period will both have 12 time points each. This is done for each project, for every month. In the end, a linear regression is applied on the first 12 months (representing the pre-conduct period) and a linear regression is applied on the last 12 months (representing the post-conduct period). The slopes and intercept of the two linear regressions are compared in order to determine if there is any difference and how large the difference is.

Another statistical technique which takes into account the passing of time is time-series [8], which are used to identify common patterns between the two periods, over time. Adapting it to the purpose of this research, would mean to identify patterns between the pre-conduct and post-conduct periods, which does not answer **RQ1**, since there is no way to know if the proportion of women has increased or decreased in the post-conduct period in comparison to the pre-conduct period. Instead, time-series would provide answers such as "in the 8^{th} month of both pre-conduct and post-conduct period, the proportion of women increased by five percent followed by a plateau which was reached in the 9^{th} month".

Besides the statistical techniques which take into account the passing of time, there are also techniques which do not take time into account. Mann Whitney U test, Wilcoxon signed rank test, and t-test are all statistical tests used to compare two or more groups (i.e. the pre-conduct group and the post-conduct group) and determine if there is a difference between them. Besides not taking into account the effect of time, they also do not take into account the impact that other additional information might have on the results. For example, some OSS projects might be more popular than others, thus there is a possibility that popularity influences how much women contribute in the project. The additional variables, such as popularity of projects, are also known as confounding variables. To ensure that the model fits well the data and that the outcome of the results are indeed correlated with the adoption of code of conduct, confounding variables need to be taken into account, thus, Mann Whitney U test, Wilcoxon signed rank test and t-test are not used in this research.

A statistical technique which does not take into account the passing of time, but takes into account confounding variables is the mixed linear regression technique. As the name suggests, it is a linear regression which can be used to determine if the proportion of women has increased or decreased in the post-conduct period, in comparison to the pre-conduct period. The only difference to RDD is the use of only one regression, applied throughout both periods, in order to determine whether an increasing or decreasing trend exists.

For the purpose of testing the null hypothesis of **RQ1**, the RDD is used for model 2 of contributorship, since the data is rich, and mixed linear regression is used for model 3 and 4 of contributorship, since the number of updates for ".all-contributorsrc", "authors" and "contributors" file is low for each period. More than 70% of the projects obtained in Section 3.8 have only one file update in the pre-conduct period and one file update in the post-conduct period. This is as expected, since maintainers of OSS projects tend to forget to update these files, which is something that the All Contributors model tried to solve by automatizing this process[26]. More about the suitability of RDD and linear regression in Sections 3.10.1 and 3.10.2. Since RDD is used for model 2 of contributorship, and linear regression for model 3 and 4 of contributorship, it is best to discuss separately the techniques used for each model.

3.10.1 Suitability of RDD for model 2 of contributorship

Model 2 of contributorship contains all the GitHub interactions, as explained in Section 3.7.1. Since these interactions represent the core of GitHub, there are more contributors and contributions, which are made more often, in comparison to model 3 and 4 of contributorship. This can also be concluded by comparing the Tables 3.9 and 3.10. Using RDD requires setting two variables: the time length period and the time length of time point.

As explained in Section 3.9.1, the length of the pre-conduct and post-conduct period needs to be the same in order to avoid skewed results. Thus, any pre-conduct or post-conduct contribution that is not in the time length period is discarded. Furthermore, the projects obtained in Section 3.8, which do not have any women's involvement in any of the pre-conduct or post-conduct periods, are removed since there is nothing to be measured for these projects, as can be seen in Table 3.13. More information about the content of the dataset can be found in Appendix B.1.

Number of projects with no women contribution	53
Number of projects with women contribution	294

Table 3.13: Overall statistics for projects which have a model 2 contribution made by women in the pre-conduct or post-conduct period

The length of a time point represents over how many days the proportion of women is measured throughout the pre-conduct and post-conduct period. Setting these variables was done according to the guidelines made by Lemieux et al. [22]. If time points have a long time length, or time length period is short, the internal validity is threatened since not enough measurements can be taken for each project. If the time length of the time points is short, the results might be skewed since there might be no women or not enough women who contribute that often in order to take measurements for each time point. If the time length period is large, there might not be enough projects which have such a lengthy contribution history for both the pre-conduct and post-conduct period. To address this problem, Lemieux et al. [22] mention using a box-plot for determining the appropriate time lengths for the periods and the time point such that there is enough data for the variable to be measured (here - proportion of women). They also insist that there is no right or wrong when choosing these values and that it all depends on the amount of data available.

Determining in which time point a contribution was done might not be so straight-forward. As mentioned in Section 3.9.1, each project has a different date when a code of conduct was adopted. Since the focus of this research is to determine how the proportions of women have evolved throughout time, before and after adopting the code of conduct, this evolution is done relative to how many days have passed since the code of conduct was adopted, or how many days before the code of conduct was adopted.

The day when the code of conduct was adopted is day 0, the day before code of conduct was adopted is day -1, and the day after code of conduct was adopted is day 1. The day when the person has contributed, relative to the date when the code of conduct was adopted, was used for determining the time-point when each person has contributed. Some examples, in which the length of time point is 30 days, might be relevant to understand the concept:

- **Example 1:** if project A has contributions made by a person on days 7, 8 and 12 (i.e. 7, 8 and 12 days after the code of conduct was introduced), their contributions will be taken into account when computing the proportion of women for project A for time-point 1.
- **Example 2:** if project A has contributions made by a a person on day 30 (i.e. 30 days after the code of conduct was introduced), their contributions will be taken into account when computing the

[26] https://allcontributors.org/docs/en/overview

proportion of women for project A for time-point 2. The reason why day 30 is on time-point 2 is that, for post-conduct time-points, day 0 is counted as being part of time-point 1. Thus, time-point 1 contains the days 0,1,..,29. Time-point 2 contains days 30,31,...,59. And so on.

- **Example 3:** if project A has contributions made by a person on days -7, -8 and -12 (i.e. 7, 8 and 12 days before the code of conduct was introduced), their contributions will be taken into account when computing the proportion of women for project A for time-point -1.

- **Example 4:** if project A has contributions made by a person on day -30 (i.e. 30 days before the code of conduct was introduced), their contributions will be taken into account when computing the proportion of women for project A for time-point -1. The reason why day -30 is on time-point -1 is that, for pre-conduct time-points, day -1 is the last day. Thus, time-point -1 contains the days -30, -29,..., -1. Time-point -2 contains days -60,-59,...,-30. And so on.

After analysing the remaining projects (i.e. the ones which have women's involvement, as shown in Table 3.13, it was concluded that the longest period is 55 months. Thus the box-plot will span over a time frame of 110 months in total (55 months for the pre-conduct and 55 months for the post-conduct period). For selecting the length of the time point, various lengths have been tried such as: 7 days, 14 days, 21 days, 30 days, 42 days, 30 days, 84 days (i.e. three months), 168 days (i.e. half a year). The conclusion when looking at all the box-plots, with different time lengths for the time points, is that anything smaller than 30 days does not have enough women to measure in each time point for each project, and everything larger than 30 days requires longer time length periods to have enough time points to apply RDD. The problem with having longer time length periods is that not enough projects have a time length period larger than one year. For example, having a time length of 1.5 months for the time points results in eight time points when using a period of one year, which might be a bit low for applying a regression and might have an impact on the results. Furthermore, having larger time lengths for the time points might loose information about how the proportion of women evolves over time. Thus, in order to have enough time points and avoid being too granular or too loose, the time length for the time points is set to 30 days. Applying RDD with time length of 30 days for the time points has been used in other papers as well [53, 63].

Figure 3.7 represents the box-plot, with the time length of 30 days for the time points, applied on the dataset obtained in Section 3.8. Each time point, found on the X-axis, is represented by a box and the color of the box represents how many projects have at least a model 2 contributorship, made by a woman, in that time point. Projects which do not have any contribution made by a woman in a certain time point, are excluded form the respective time point. This explains why the boxes have different colors, since not all projects have a contribution made by a woman in every time point. Each period is split into 55 time points. Each box measures the proportion of women who have contributed in the respective time point, for each project, as explained in Section 3.9.1. The proportion of women is found on the Y-axis. Negative values on the X-axis represent time-points from the pre-conduct period, while positive values represent time-points from the post-conduct period. The time-point associated with the introduction of code of conduct has value 1 on the X-axis. Everything mentioned until now is as recommended by Thomas Lemieux et al. in their RDD guidelines [22].

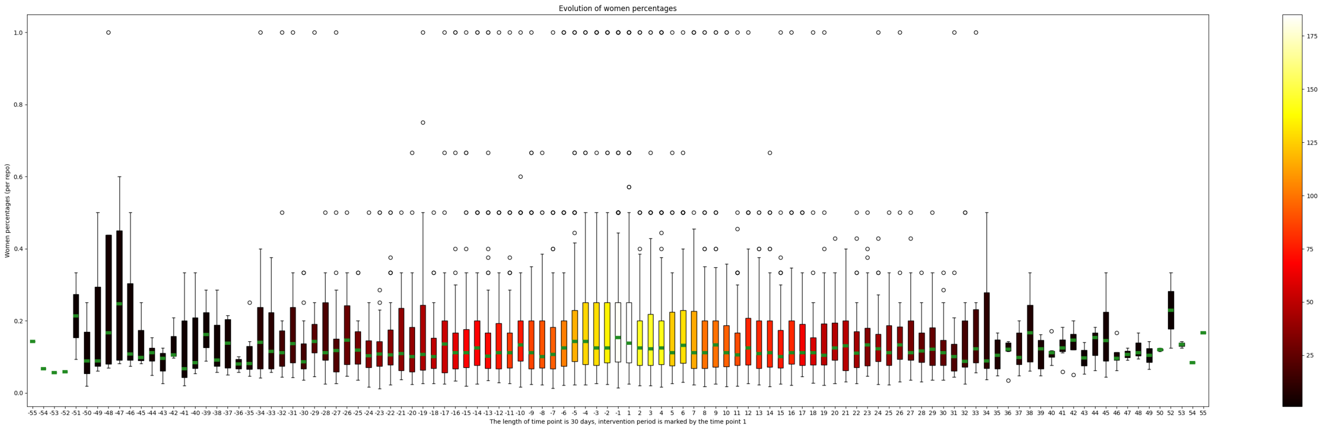


Figure 3.7: Plot showing, for every 30 days, the proportion of women who have contributed in each project. The introduction of code of conduct is represented by time-point 1. Projects which do not have any contribution made by a woman in a certain time point, are excluded from the respective time point.

Analyzing the colors of each time point from the box-plot from Figure 3.7, concludes that the number of projects which have women contribution starts to decrease after the time-point 12 and before time-point -12. This is further explained by the fact that out of 294 projects, only 160 have a time length period longer than 360 days. Because of this decrease, it was decided to set the time length period across all projects to 360 days. Projects which have either a time length period smaller than 360 days, or no contribution made by women in any of the 12 time points before or after the code of conduct was adopted, are removed from the dataset. The number of projects removed is 145, and the number of projects that remained is 149. These numbers can be seen in Table 3.14 as well. More information about the content of the dataset can be found in Appendix B.2. The 360 days period length is referenced throughout this thesis as the "defined period length". Furthermore, for each time-point from the defined period length, there are at least 80 projects which have contributions made by women. Thus, there are enough projects to fit a RDD model [22]. No increasing or decreasing trend, with respect to proportion of women, can be seen in Figure 3.7. Comparing both periods, the number of projects which have women contribution stays the same, no significant increase in projects which have women contribution can be seen in the post-conduct period, suggesting that code of conduct has little to no effect. This conclusion is further strengthen when looking at the results of the RDD for model 2 in Section 4.1.1.

Number of projects with no women contribution in the defined period length or with period length smaller than 360 days	145
Number of projects with women contribution in the defined period length and with period length larger or equal to 360 days	149

Table 3.14: Overall statistics for projects which have a model 2 contribution made by women in the defined period length and with period length larger or equal to 360 days

Figure 3.8 contains all the remaining projects, after removing the ones which do not have any model 2 contribution made by women in the defined period length or which have a period length smaller than 360 days. In Figure 3.8, also the projects which do not have any contribution made by a woman in a certain time point, are still kept in the respective time point. This explains why the boxes have the same colors, since all 149 projects are kept at every time point. The RDD model will be applied on all the information shown in the aforementioned figure. Similar to before, no increasing or decreasing trend, with respect to proportion of women, can be seen.

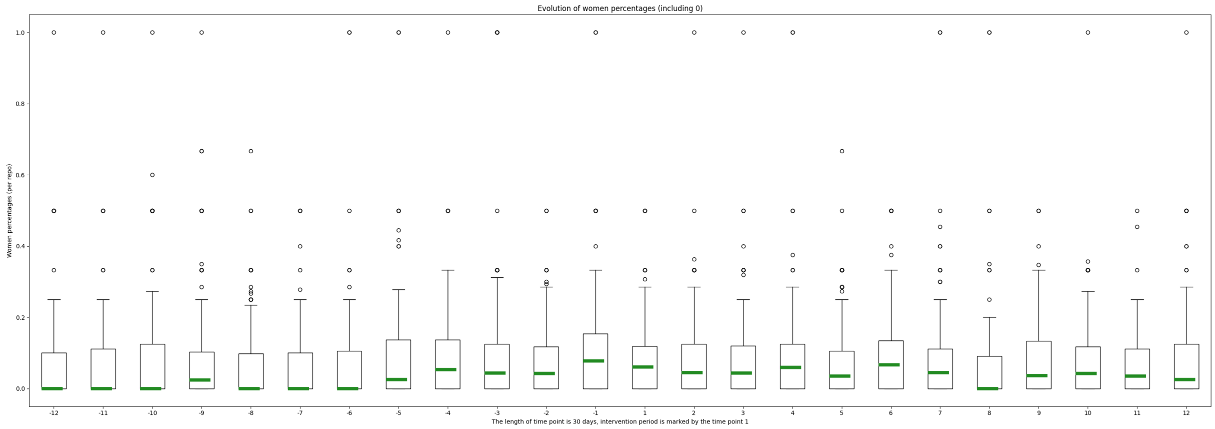


Figure 3.8: Plot showing, for every 30 days, the proportion of women who have contributed in each project. The introduction of code of conduct is represented by time-point 1. Projects which do not have any contribution made by a woman in a certain time point, are kept in the respective time point.

3.10.2 Suitability of mixed linear regression for model 3 and 4 of contributorship

For model 3 and 4 of contributorship, the same time length period, as the one for RDD, is used. The reason is to ensure that the contribution is measured for the same time frame across the two statistical techniques. To further remove threats and biases, only the 149 projects, found in Table 3.14, are used for the mixed linear regression. In this way, the data stays the same across the two statistical techniques.

The reason why RDD was not used for model 3 and 4 of contributorship as well, is that 96 out of the 149 projects have only one update for the ".all-contributorsrc", "authors" and "contributors" file in the pre-conduct and post-conduct period. An RDD requires two linear regressions, one for the pre-conduct period and one for the post-conduct period, and each of the linear regression requires at least two points, which are not present in the dataset. Thus, the only statistical technique which determines whether there is an increase or decrease in the proportion of women between the two periods, while taking into account for confounding variables, is the mixed linear regression. The mixed linear regression was preferred instead of the linear regression, since the former one takes into account for confounding variables, as explained in Section 3.10. In case a project has multiple file updates, for the pre-conduct period, the file-update closest to the introduction of code of conduct is kept, and for the post-conduct period, the file-update furthest from the introduction of code of conduct is kept. Similar to RDD, any file updates after the pandemic started, 01-March-2021, are not taken into account, as well as file updates that happen outside of the defined period. The latter one was done to ensure that both model 2 and model 3 and 4 of contributorship are analyzed over the same time frame. Since each project has one contributors list in the pre-conduct period, and one contributors list in the post-conduct period, each project will have only two women proportion values. These two values for each project are referenced throughout this thesis as time-points as well, to have the same consistency as the RDD. As explained in Section 3.7.2, the contributors name were manually extracted from the lists. The proportion is computed by counting how many women there are in a list and dividing it by the total people from the list. This is done separately for the pre-conduct and post-conduct list, for each project. As shown in Table 3.10, there are more than enough contributors to compute proportions for the 149 projects and apply mixed linear regression.

3.11 Statistical techniques for RQ2

To test the null hypothesis mentioned in Section 3.9.2, one or more suitable statistic technique needs to be selected. Based on the statistical techniques, statistical models are constructed to test the null hypothesis.

While statistical techniques such as Mann Whitney U test, Wilcoxon signed rank test, and t-test can be used to determine if there is a difference between women who contributed in the pre-conduct period in comparison to women who contributed in the post-conduct period, with respect to time women remained contributors in a project, they do not show how the presence of code of conduct influences women to stay or leave the project in the long run. For example, it does not show whether women are more likely to stay in the project for a time length longer than, for example, three months, or longer than one year, or longer than any other time period. Such information cannot be obtained from the aforementioned statistical techniques since they are not designed to estimate how likely a person is to stay in a project at various time lengths. The aforementioned tests can say whether the two periods are different or not. For example, it might be that with, or without a code of conduct, women have the same probability of staying three months in a project, but it might be that in the presence of a code of conduct, women are more likely to stay a year in the project, in comparison to when there is no code of conduct. For obtaining these insights, and for better answering **RQ2**, different statistical techniques should be used.

Determining how long contributors remain in the pre-conduct period, in comparison to the contributors from the post-conduct period, can be interpreted as a measure of time until an event happens for the respective contributor, where the event here is the person leaving the project. A common statistical technique for measuring time until event, also known as survival time, is the survival analysis [25]. As the name suggests, this technique estimates a survival probability function for the pre-conduct period and a survival probability function for the post-conduct period. The advantage of having such probability functions is the fact that they can be plotted in order to visually inspect how the code of conduct has influenced the time a contributor remains in a project. Since the dataset contains only two groups of people: women who have contributed in the pre-conduct period, and women who have contributed in the post-conduct period; the Kaplan–Meier estimator is used to determine the survival probability functions [25]. Survival analysis is used especially in the medical field for determining how a certain medication influences the chances of survival for patients which suffer from a certain disease. Furthermore, determining if contributors remain in a project for a longer duration of time after code of conduct was adopted, can also be interpreted as measuring the effectiveness of a treatment, where the treatment here is the presence of code of conduct.

While the survival analysis is great for visual inspection, by plotting the survival analysis curves, it does not tell whether the two groups are different. For determining whether the two groups are different, a log-rank test is applied on the survival analysis functions [25].

The downside when using survival analysis is that it does not take into account the impact that other additional information might have on the results. For performing a survival analysis, while taking into account for confounds, a different statistical technique is commonly used, named the Cox proportional-hazards [25]. The Cox proportional-hazards is an extension of the survival analysis. The advantage of using this technique is that it determines the impact that other additional information might have on the survival analysis results [25].

3.11.1 Suitability of survival analysis, log-rank test and Cox proportional-hazards

For answering **RQ2**, the survival analysis, with Kaplan–Meier estimator, is chosen to visualize and compare the survival probability functions between the two periods. Log-rank test is used to compare the survival curves drawn by the survival probability functions. To determine whether code of conduct has an impact on the time women remain contributors in projects, the Cox proportional-hazards is used. Since the Log-rank test and Cox proportional-hazards are based on the results obtained from the survival analysis, the suitability of survival analysis is checked in this section.

As explained in Section 3.9.2, only model 2 contributions from both pre-conduct and post-conduct period are used. For answering **RQ2**, the number of repositories which have women interactions are determined, together with the number of women who contributed only in the pre-conduct period or only in the post-conduct period. If there are few repositories, the results might be threaten since they cannot be generalized [39]. If there are few women for the survival analysis, the internal validity might be threaten as well [39]. From the model 2 of contributors dataset, obtained in Section 3.8, only the women who have contributed only in one of the periods, are kept. The repositories which do not have any women contribution, are removed from the dataset. The number of projects and women that are kept in the dataset, are available in Table 4.1.

Number of projects with women who have contributed in only one period	347
Number of women who have contributed only in pre-conduct period	3,826
Number of women who have contributed only in post-conduct period	4,123

Table 3.15: Number of projects which have a model 2 contribution made by women only in one of the periods. Number of women who have contributed in only one of the periods.

For comparing the time women remained in the pre-conduct period, with the time women remained in the post-conduct period, a survival probability function is used for each of the periods, using the Kaplan–Meier estimator. The survival probability function, for the pre-conduct period, is determined based on 3,826 women who have contributed only in the pre-conduct period, and the survival probability function, for the post-conduct period, is determined based on 4,123 women who have contributed only in the post-conduct period. Taking into account the number of women analysed for each of the periods, the survival probability should have enough data points for determining each of the survival probability functions.

3.12 Statistical models used for RQ1 and RQ2

In this section the statistical models needed for testing the null hypotheses for **RQ1** and **RQ2**, are presented. The research questions that we are trying to answer in this thesis are:

RQ1:*Do projects have a (higher) increase in the proportion of women who contribute in the project, in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?*

RQ2:*Do women remain contributors over a longer duration of time in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?*

As mentioned in Section 3.9, the following null hypotheses are used for answering **RQ1** and **RQ2**:

H0 for RQ1:*Adopting a code of conduct does not have any influence on the proportion of women who contribute in a project*

H0 for RQ2:*Adopting a code of conduct does not have any influence on the time women remain contributors in a project*

In Section 3.12.1, the statistical models used for answering **RQ1** are presented, and in Section 3.12.2, the statistical models used for answering **RQ2** are presented.

3.12.1 Statistical models for RQ1

In this section the statistical models used to fit the RDD and mixed linear regression techniques, to analyze **RQ1**, are described. The features of interest, used to minimize confounds, are discussed in this section as well. Despite having two different techniques, the features of interest remain the same for both statistical models, since they are applied on the same dataset, thus, there is no need to discuss them separately.

The following features are of interest:

- **Project name:** the name of the project, to which the contributions belong. The name is used to identify the project in GitHub;
- **Programming language:** the primary language used in the project, as reported by GitHub. The reason for taking the language into account is that some programming languages are more popular or easier to use, for example Python, in comparison to other, such as Haskell. This variation in preferences needs to be taken into account when measuring contribution;
- **Number of code of conduct updates:** in the paper by Singh et al. [48], women mentioned that code of conduct are more efficient if they are perceived as being an important part of the project. One way to measure it is by the number of times the code of conduct file was updated and maintained, since new versions of code of conduct are released constantly;

- **Code of conduct is link:** another way to measure how important is a code of conduct in a project, is to analyze the content of the code of conduct file. After looking through all 149 projects, it was concluded that some code of conduct files just contain a link to a popular code of conduct, in which the members of the project can read more about what a code of conduct entails. Because of this difference, in content and approach, it was decided to take such variations into account as well;
- **Code of conduct is custom:** another way to measure how important is a code of conduct in a project, some projects adopt a standard code of conduct, such as the one made by: Contributor Covenant, Mozilla, Python, Django; but some projects also adopt a custom code of conduct, made specifically for their project community. Thus, the code of conduct content for each of the 149 projects was manually analyzed to determine if it is custom or not;
- **Code of conduct has contact:** in the paper by Singh et al. [48], women mentioned that it is important to have someone to contact in case the rules of the code of conduct have been broken. Thus, the code of conduct content for each of the 149 projects was manually analyzed to determine if contact information is provided to report violations of code of conduct rules;
- **Number of stars:** a study by Balali et al. [3], which interviewed women and men to find the reasons that motivate them to contribute in OSS, concluded that men choose an OSS project to contribute on, mostly on popularity, because they feel that it is their duty and see it more as a job, while women choose a project because they have a real interest in it or the community around it. Women stated that ”for men it’s more their job to contribute to OSS, but women want to do it because they find it exciting” [3]. Since measuring how exciting a project is, is highly subjective and nearly impossible to do without adding more threats and assumption to the research, it was decided to measure the popularity of the project instead. The intuition is, that the more popular a project is, the more people would like to contribute on it, and stay longer, thus, the proportion of women might be influenced by the popularity of the project, and thus, needs to be accounted for. Number of stars[27] is a feature on GitHub in which people can show appreciation towards a project. The more stars a project has, the more it has been seen or used by people. The number of stars is correlated with higher contribution and is a good metric to measure the popularity of a project [7];
- **Number of forks:** similar to number of stars, number of forks is another metric which might be used to measure popularity. The number of forks that a project has, represents the number of times the project has been cloned until now. The more clones, the more interest it generates or the more influential it is in the OSS community. As mentioned in Section 3.6.3, in the dataset there were more than 200 clones for the TensorFlow project. Number of forks have been used in studies to reflect the popularity of a project [64];
- **Gender of repository owner:** studies have emphasized the importance that women role models have on other women in the community, helping them feel more welcomed [56]. Having a woman as a role model in the project helps with the peer parity, making them feel more confident in their work and reduces the “impostor syndrome” that women are dealing with. For this study, the role models are represented by the repository owner. Furthermore, the intuition is that a repository owned by a woman might have a lower tolerance to discrimination towards women.
- **Type code of conduct:** the type of conduct is represented by the code of conduct organization which created the code of conduct, such as Contributor Covenant, Mozilla, Python. despite studies showing that all code of conduct types promote the same values and enforce the same rules, it was decided to take this variation into account as well since people might perceive them differently.

The dependent variable for both RDD and the mixed linear regression is the proportion of women who have contributed in a certain time-point. The number of code of conduct updates, number of stars, watchers and forks are normalised, using min-max normalization[28], across the entire dataset. For the other features, no normalization is required since the dependent feature is expressed in percentages, with a range between 0 and 1, while the other independent features are categorical. As mentioned in Section 3.10.1, projects which

[27] https://github.blog/2012-08-06-notifications-stars/

[28] https://www.codecademy.com/articles/normalization

do not have any contribution made by women are removed in order to reduce bias. To remove threats due to inconsistency of the data and variances across projects, both statistical models are applied on the same projects and their contribution is measured across the same time period.

The differences between the two statistical models is that RDD takes into account the passage of time. To achieve this, for RDD, three additional features are used: *time*, *time after intervention* and *intervention*. The first feature is used to measure the passing of time for the pre-conduct period, and has numeric values, from 1 to 12, which are assigned to time points in the following way: the first time point, i.e. 12 months before the code of conduct was adopted in the project, has value 1, the time point just before intervention has value 12. For the time points from the post-conduct period, the value is set to 0. The variable *time after intervention*, is used to model the post-conduct period, in a similar way as for *time*. Variable *intervention* is set to 0 for all the time-points that belong to the pre-conduct period, and to 1 for the ones which belong to the post-conduct period. This process is done for each project out of the 149 projects, found in 3.14. Looking at the significance of these three variables, determines whether the code of conduct has an impact on the proportion of women in the post-conduct period, in comparison to the pre-conduct period. In case there is an impact, looking at the coefficients of the significant variables, determines how large the impact was on the proportion of women. This entire procedure for the RDD is standard and has been used in other research papers as wel [63].

Another difference is that the mixed linear regression model has only the variable *intervention*, in comparison to RDD. The *intervention* is set to 0 for pre-conduct period, and to 1 for the post-conduct period. In this way, the proportion of women from the pre-conduct period can be compared with the proportion of women from the post-conduct period, to determine if there is any statistical difference between them.

Implementing these models was done in R[29], using the **lmerTest** package. To take into account for the variability and differences across projects, such as the way the community interacts and contributes, the independent features *project name*, *programming language* and *type code of conduct* are modeled as random effects. Another important aspect is to remove covariance[30]. Colinearity between variables might lead to wrong results, thus, it is recommended to remove them. For this, a variance inflation factor (VIF) was computed for each feature, and features which have a VIF higher than 5 should be removed [45]. All features have a VIF smaller than 3, for both models.

To determine how well the models fit the dataset, the marginal, R^2_m, and conditional, R^2_c, r-squared values are used [34]. The former one represents the goodness of fit when not taking into account the random variables, while the latter one represents the goodness of fit when the random effects are taken into account.

3.12.2 Statistical models for RQ2

In this section the statistical models used to fit the survival analysis, with Kaplan–Meier estimator, as well as the Cox proportional-hazards techniques, are described. Since the survival analysis cannot take into account for additional information, the survival analysis, and the Log-rank test, contain only the dependent variables for each of the periods.

The Cox proportional-hazards technique can support additional information, thus the features of interested used in the statistical model used for this technique, are the followings:

- **Intervention:** is set to 0 for all the women who have contributed in the pre-conduct period, and to 1 for the women who have contributed in the post-conduct period. This variable is used to determine the impact that code of conduct has on time women remain contributors in a project.
- **Project name:** the name of the project, to which the contributions belong. The name is used to identify the project in GitHub;
- **Programming language:** the primary language used in the project, as reported by GitHub. The reason for taking the language into account is that some programming languages are more popular or easier to use, for example Python, in comparison to other, such as Haskell. This variation in preferences needs to be taken into account when measuring contribution;

[29] https://www.r-project.org/

[30] https://en.wikipedia.org/wiki/Covariance

- **Number of code of conduct updates:** in the paper by Singh et al. [48], women mentioned that code of conduct are more efficient if they are perceived as being an important part of the project. One way to measure it is by the number of times the code of conduct file was updated and maintained, since new versions of code of conduct are released constantly;
- **Code of conduct is link:** another way to measure how important is a code of conduct in a project, is to analyze the content of the code of conduct file. After looking through all 347 projects, it was concluded that some code of conduct files just contain a link to a popular code of conduct, in which the members of the project can read more about what a code of conduct entails. Because of this difference, in content and approach, it was decided to take such variations into account as well;
- **Code of conduct is custom:** another way to measure how important is a code of conduct in a project, some projects adopt a standard code of conduct, such as the one made by: Contributor Covenant, Mozilla, Python, Django; but some projects also adopt a custom code of conduct, made specifically for their project community. Thus, the code of conduct content for each of the 347 projects was manually analyzed to determine if it is custom or not;
- **Code of conduct has contact:** in the paper by Singh et al. [48], women mentioned that it is important to have someone to contact in case the rules of the code of conduct have been broken. Thus, the code of conduct content for each of the 347 projects was manually analyzed to determine if contact information is provided to report violations of code of conduct rules;
- **Number of stars:** a study by Balali et al. [3], which interviewed women and men to find the reasons that motivate them to contribute in OSS, concluded that men choose an OSS project to contribute on, mostly on popularity, because they feel that it is their duty and see it more as a job, while women choose a project because they have a real interest in it or the community around it. Women stated that "for men it's more their job to contribute to OSS, but women want to do it because they find it exciting" [3]. Since measuring how exciting a project is, is highly subjective and nearly impossible to do without adding more threats and assumption to the research, it was decided to measure the popularity of the project instead. The intuition is, that the more popular a project is, the more people would like to contribute on it, and stay longer, thus, the time women remain contributors might be influenced by the popularity of the project, and thus, needs to be accounted for. Number of stars[31] is a feature on GitHub in which people can show appreciation towards a project. The more stars a project has, the more it has been seen or used by people. The number of stars is correlated with higher contribution and is a good metric to measure the popularity of a project [7];
- **Number of forks:** similar to number of stars, number of forks is another metric which might be used to measure popularity. The number of forks that a project has, represents the number of times the project has been cloned until now. The more clones, the more interest it generates or the more influential it is in the OSS community. As mentioned in Section 3.6.3, in the dataset there were more than 200 clones for the TensorFlow project. Number of forks have been used in studies to reflect the popularity of a project [64];
- **Gender of repository owner:** studies have emphasized the importance that women role models have on other women in the community, helping them feel more welcomed [56]. Having a woman as a role model in the project helps with the peer parity, making them feel more confident in their work and reduces the "impostor syndrome" that women are dealing with. For this study, the role models are represented by the repository owner. Furthermore, the intuition is that a repository owned by a woman might have a lower tolerance to discrimination towards women.
- **Type code of conduct:** the type of conduct is represented by the code of conduct organization which created the code of conduct, such as Contributor Covenant, Mozilla, Python. despite studies showing that all code of conduct types promote the same values and enforce the same rules, it was decided to take this variation into account as well since people might perceive them differently.

The dependent variable for both the survival analysis and the Cox proportional-hazards model is the time

[31] https://github.blog/2012-08-06-notifications-stars/

women remain contributors in a project, adjusted to the length of the pre-conduct or post-conduct period of each project, and adjusted to the time when the contributor has joined the project. The number of code of conduct updates, number of stars, watchers and forks are normalised, using min-max normalization[32], across the entire dataset. For the other features, no normalization is required since the dependent feature is expressed in percentages, with a range between 0 and 1, while the other independent features are categorical.

To determine how well the models fit the dataset, the marginal, R_m^2, and conditional, R_c^2, r-squared values are used [34]. The former one represents the goodness of fit when not taking into account the random variables, while the latter one represents the goodness of fit when the random effects are taken into account.

[32] https://www.codecademy.com/articles/normalization

Chapter 4

Results

In this chapter, the results obtained from the statistical models described in Section 3.12 are presented. The results for **RQ1**, together with a discussion of results for **RQ1**, are presented in Section 4.1 and the results for **RQ2**, together with a discussion of results for **RQ2**, are presented in Section 4.2. Threats to validity for **RQ1** are presented in Section 4.1.4, and the threats to validity for **RQ2** are presented in Section 4.2.2. An overall discussion, taking into account the results for **RQ1** and **RQ2** is presented in Section 4.3. In Section 4.4, the threats to validity from combining the aforementioned answers are presented.

4.1 Results for RQ1

In this section, the results for the two models used to answer RQ1 are presented. The results from the RDD, applied only on model 2 of contributorship, are presented in Section 4.1.1, while the results from the mixed linear regression, applied only on model 3 and 4 of contributorship, are presented in Section 4.1.2. An answer to **RQ1** is provided in Section 4.1.3. The threats to validity, for **RQ1**, are presented in Section 4.1.4.

4.1.1 Results for model 2 of contributorship

The RDD model used for determining the impact that code of conduct has on the proportion of women, is the one described in Section 3.12.1. In Table 4.1 the results of the RDD model are presented, containing the significance and impact for each of the features used, as well as the variance explained by each feature. Analyzing Table 4.2, it can be concluded the variable *time* has a significant impact, with a positive coefficient, meaning that the proportion of women was increasing in the pre-conduct period. The variable *intervention* has also a significant impact, with a positive coefficient, meaning that in the month when the code of conduct was adopted, the proportion of women has increased by 0.01627%. The last variable used to model time is *time after intervention*, which has no significant impact meaning that there is not enough evidence to suggest that there is an increasing or decreasing trend, with respect to proportion of women, in the post-conduct period.

Table 4.1: significance and impact of the features used in RDD to model the impact that code of conduct has on proportion of women, for model 2 contributorship

Features	**Coeffs.**	**Sum sq.**
Intercept	8.834e-02	4.375e-02
Time**	1.945e-03	4.221e+03
Intervention**	1.627e-02	9.770e+02
Time after intervention	-4.402e-04	4.224e+03
Number code of conduct updates	-1.328e-02	1.804e+02
Code of conduct is link	4.014e-02	2.091e-02
Code of conduct has contact	1.634e-02	1.114e-02
Code of conduct is custom	2.798e-02	1.897e-02
Number of stars	-5.384e-02	1.373e+02
Number of forks	7.641e-02	1.358e+02
Gender of repository owner - Men**	-1.442e-01	1.222e+02
Gender of repository owner - No gender**	-1.297e-01	1.208e+02

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Out of all the independent variables, only the gender of the repository owner has a significant impact on proportion of women. As the literature suggested [56], having a male role model, or no role model, has a negative impact on the proportion of women. Over a period of one year, in projects with male or unidentified owners, the proportion of women will be 2% smaller than if the projects had women owners. The model fits well, having a marginal score of 0.126 and a conditional score of 0.572. Thus, when taking into account the variability introduced by the differences between projects, the model fits well the proportion of women dataset.

An immediate positive jump in the proportion of women can be observed when the code of conduct is adopted. However, there is no observable increasing or decreasing trend, with respect to the proportion of women, in the period after the code of conduct was adopted.

With respect to rejecting or not the null hypothesis, the null hypothesis is rejected. Code of conduct has an influence on the proportion of women, but only on a limited amount of time. In the long run (i.e. after the first month of adopting code of conduct), code of conduct does not have any impact on the proportion of women.

4.1.2 Results for model 3 and 4 of contributorship

For determining the impact that code of conduct has on model 3 and 4 of contributorship, a mixed linear regression was used, as described in Section 3.12.1. In Table 4.2 the results of the mixed linear regression model are presented, containing the significance and impact for each of the features used, as well as the variance explained by each feature. Analyzing Table 4.2, it can be concluded the variable *intervention* has no significant impact meaning that there is not enough evidence to suggest that the proportion of women has increased or decreased after the code of conduct was adopted.

Table 4.2: significance and impact of the features used in the mixed linear regression to model the impact that code of conduct has on proportion of women, for model 3 and 4 of contributorship

Features	**Coeffs.**	**Sum sq.**
Intercept	1.283e-02	1.252e-01
Intervention	1.349e-02	1.191e-02
Number code of conduct updates	-8.031e-03	1.226e-01
Code of conduct is link	3.160e-02	5.123e-02
Code of conduct has contact	-3.309e-02	3.098e-02
Code of conduct is custom	4.328e-02	4.184e-02
Number of stars	-1.805e-01	1.712e-01
Number of forks	5.444e-02	2.955e-01
Gender of repository owner - Men	-2.588e-05	9.243e-02
Gender of repository owner - No gender	3.531e-02	8.820e-02

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Out of all the independent variables, none has a significant impact on the proportion of women. The model fits well, having a marginal score of 0.119 and a conditional score of 0.548. Thus, when taking into account the variability introduced by the differences between projects, the model fits well the proportion of women dataset.

No increase or decrease in proportion of women is observed in the period after the code of conduct was adopted, in comparison to the period before code of conduct was adopted.

With respect to rejecting or not the null hypothesis, the null hypothesis cannot be rejected since the variable *intervention* does not have any significance on the proportion of women. Thus, code of conduct does not have an influence on the proportion of women.

4.1.3 Discussion for RQ1

In order to answer **RQ1**, *Do projects have a (higher) increase in the proportion of women who contribute in the project, in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?*, the results obtained from RDD and mixed linear regression are analyzed.

Despite RDD rejecting the null hypothesis, while the mixed linear regression does not, it is important to understand that these two models are applied on two different datasets: RDD is applied on the model 2 of contributorship, while the mixed linear regression is applied on model 3 and 4 of contributorship. There is a difference in terms of available information, RDD applies two linear regression, each one over 12 time points, while mixed linear regression applies only one linear regression over two time points. Because of the differences, with respect to the number of time points, RDD can measure the impact that code of conduct has on proportion of women, right after the code of conduct was adopted, as well as the impact that code of conduct has after the first month in which the code of conduct was adopted. The mixed linear regression model fails to capture the initial spike in proportion of women, since none of the 149 projects have an update for contributors file in the month when the code of conduct was adopted, thus the mixed linear regression model fails to capture the initial spike in proportion of women.

If the month before the code of conduct was adopted and the month after it, are excluded, the null hypothesis cannot also be rejected by RDD. In this case, both models reveal the same thing: code of conduct does not have any impact on the proportion of women. As seen in Figure 3.7, the month in which the code of conduct was adopted and the month before that, have the highest number of projects which have contributions made by women. The increased activity might explain the initial jump in proportion of women.

Since the focus of this thesis is on analyzing the pre-conduct and post-conduct period, and especially the long term effects that code of conduct has on women's involvement, it was decided to exclude the initial jump in proportion of women. As mentioned before, a code of conduct might take time in order to be adopted by a community, people might leave, new people might join. A community might not become more friendly to women over-night, instead, it is a process that takes time. Furthermore, to evaluate the effect of a treatment or change (e.g., a new drug on a disease, a new rule on a community) a longitudinal study is suited [9], in which data is collected and analyzed over a long period of time. Taking into account the results from the RDD and mixed linear regression, excluding the initial jump in proportion of women, there is no significant evidence that adopting a code of conduct has any impact, on the proportion of women who contribute in the project. Since there is no statistical evidence to suggest that code of conduct increase or decreases the proportion of women in the post-conduct period, no estimations can be made to determine the impact that code of conduct has on women, after the month in which the code of conduct was adopted. Thus, the answer for **RQ1** is the following:

Code of conduct does not lead to any observable increase or decrease, in proportion of women who contribute in a project, in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted. Code of conduct does not have any observable impact on the proportion of women who contribute in a project.

4.1.4 Threats to validity for RQ1

Internal validity: to reduce the threat of confounds, which might influence results, fixed and random factors were taking into account in the statistical models. Features based on which there is scientific proof that might influence the involvement of women, were taken into account. However, there might be other, unknown, factors which might influence the results. To address this issue, the conditional and marginal scores were used to ensure that the models fit the data well.

Construct validity: the choice of the defined length period and of the time length of the time points might also influence the results. In order to reduce this, a visual analysis was done to determine which time lengths have the most data to analyze, as suggested in the RDD guidelines by Lemieux et al. [22]. The intervention point (i.e. the point which separates the pre-conduct period from the post-conduct period) might also influence the results. To reduce the threat, the date when the code of conduct was added for the first time in the project, is considered as the intervention point. There are research papers [63] which consider a buffer time, representing a gap between the pre-conduct and post-conduct period. It was decided to not use any buffer zone, so the intervention point is not influenced by any decisions made in this research.

4.2 Results for RQ2

In this section, the results for the survival analysis, Log-rank test and Cox proportional-hazards model, used to answer RQ2, are presented. The results apply only for model 2 of contributorship, as explained in Section 3.9.2. An answer to **RQ2** is provided in Section 4.2.1. The threats to validity, for **RQ2**, are presented in Section 4.2.2.

The graph obtained from plotting the survival probability function for each of the periods, across all 347 projects, is presented in Figure 4.1. The Y axis represents the probability that a woman remains in a project for X amount of time, where X represents a value on the X axis. The X axis represents the time woman remains in a project, adjusted to the time when she has joined the project and the time length of the pre-conduct or post-conduct period, as discussed in Section 3.9.2. The plot of the survival probability function for the pre-conduct period is in green, while the plot of the survival probability function for the post-conduct period is in orange. Analyzing the chart, presented in Figure 4.1, it was concluded that both survival probability functions follow the exact curve and that there is no difference between the pre-conduct and post-conduct period, with respect to time women remain contributors in a project. The Log-rank test, which has a p-value of 0.5, also confirms that the null hypothesis cannot be rejected, concluding that there is no difference between the pre-conduct and post-conduct period, with respect to time women remain contributors in a project.

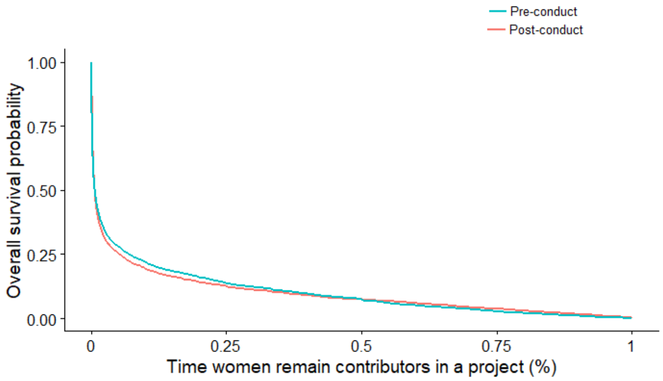


Figure 4.1: Survival analysis showing the survival probability curves for time women remain contributors in a project, in the pre-conduct and post-conduct period. The time women remain contributors in a project is adjusted taking into account the time length of the pre-conduct and post-conduct for each project, as well as the time when each woman joined the project. Only model 2 of contributorship is taken into account.

The Cox proportional-hazards model, used for determining the impact that code of conduct has on time women remain contributors in a project, is the one described in Section 3.12.2. In Table 4.3 the results of the Cox proportional-hazards model are presented, containing the significance and impact for each of the features used. Analyzing Table 4.3, it can be concluded the variable *intervention*, representing the adoption of code of conduct, has no impact on time women remain contributors in a project. Since code of conduct does not have any impact on time women remain contributors in a project, the null hypothesis cannot be rejected, concluding that there is no difference between the pre-conduct and post-conduct period, with respect to time women remain contributors in a project.

Table 4.3: significance and impact of the features used in Cox proportional-hazards to model the impact that code of conduct has on time women remain contributors in a project, for model 2 contributorship

Features	**Coeffs.**
Intervention	0.03390
Number code of conduct updates	0.01882
Code of conduct is link	0.780
Code of conduct has contact	-2.329e-03
Code of conduct is custom	5.372e-02
Number of stars	0.405
Number of forks	0.337
Gender of repository owner - Men**	0.049
Gender of repository owner - No gender**	0.0297

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Out of all the independent variables, from Table 4.3, only the gender of the repository owner has a significant impact on proportion of women. As the literature suggested, having a male role model, or no role model, has a negative impact on time women remain contributors in a project. The model fits well, having a marginal score of 0.11 and a conditional score of 0.532. Thus, when taking into account the variability introduced by the differences between projects, the model fits well the proportion of women dataset.

Code of conduct does not have any observable impact on time women remain contributors in a project. There is no observable difference between the pre-conduct and post-conduct period, with respect to time women remain contributors in a project.

4.2.1 Discussion for RQ2

In order to answer **RQ2**, *Do women remain contributors over a longer duration of time in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted?*, the results obtained from survival analysis, log-rank test and Cox proportional-hazards model are analyzed.

None of the results obtained rejects the null hypothesis, concluding that adopting a code of conduct does not have any impact on the time women remain contributors in a project. The only additional information that has an impact on the time women remain contributors in a project, is the gender of the repository owner. Projects in which the gender of the repository owner could not be defined, or in which the repository is owned by a man, had a negative impact on women, the expected hazard, on time women remain in a project, is 1.05 times higher (i.e., $\exp(0.049) = 1.05$) in projects where men are owners, and 1.04 times higher (i.e., $\exp(0.0297) = 1.04$) in projects where the gender of the owner is not identified, in comparison to projects where women are owners.

Analyzing the survival analysis, Figure 4.1, code of conduct does not have an impact in the long run, nor in the short run, when it comes to time women remain contributors in a project. In essence, the chances that a woman is remains a contributor for any certain amount of time, is the same in both periods.

The results obtained from the Cox proportional-hazards model, conclude that there is no statistical evidence to suggest that the adoption of code of conduct increases or decreases the time women remain contributors in a project. Because of the lack of statistical evidence, no estimations can be made to determine the impact that code of conduct has on the time women remain contributors in a project. Thus, the answer for **RQ2** is the following:

Code of conduct does not lead to any observable increase or decrease, in the time women remain contributors in a project, in the period after the code of conduct was adopted, in comparison to the period before the code of conduct was adopted. Code of conduct does not have any observable impact on the time women remain contributors.

4.2.2 Threats to validity for RQ2

Internal validity: the presence of confounding factors might still represent a threat. The survival analysis and Log-rank test do not take into account for confounding factors, which might influence the results and conclusions. To reduce the threat of confounds, the Cox proportional-hazards model is used, which takes into account the fixed and random factors. Only features based on which there is scientific proof that might influence the involvement of women, were taken into account. However, there might be other, unknown, factors which might influence the results, for which there is no scientific proof that they might influence the time women remain contributors. To address this issue, the conditional and marginal scores were used to ensure that the models fit the data well.

Construct validity: women might have joined a project before the first contribution was made on GitHub. Contributions and interactions outside of GitHub, such as slack, emails or other online communication platforms were not included in this analysis, because some information is not publicly available, and the communication platforms are very diverse, each one having different APIs and limitations to extract the data.

Women might have left the project after the last contribution was made on GitHub. To minimise this, the date when the person has left the project to be extended by a certain number of days, since the person might still be active in the project despite making no model 2 contributions. These added days are computed for each person in each project, by taking the longest time length between two consecutive contributions the respective person made in the respective project.

4.3 Discussion

The focus of this thesis was to determine the impact that code of conduct has on women's involvement. Since the focus of this thesis is rather broad, it was narrowed to only two research questions, **RQ1** and **RQ2**, as an attempt to focus only on certain aspects of women's involvement. **RQ1** is focused on proportion of women who contribute in OSS projects, while **RQ2** is focused on how much time women remain contributors in

OSS projects. After analyzing the data obtained in Chapter 3, and performing statistical tests in Sections 4.1 and 4.2, for answering **RQ1** and **RQ2** respectively, the conclusions are that code of conduct might have a short term impact on the proportion of women who contribute in OSS projects, but there is no statistical significant evidence that code of conduct might have, in the long run, any impact on the time women remain contributors in OSS projects.

A possible explanation for the lack of impact that code of conduct has, in the long run, on proportion of women, and on time women remain contributors in OSS projects, is that adding a code of conduct file in the OSS project, might not mean that the community embraced the rules described in the code of conduct, or it might be that the community does not respect them. As shown in the paper by Singh et al. [48], besides having a code of conduct in OSS projects, it is important how the code of conduct is enforced in OSS projects. Determining whether a code of conduct is enforced in an OSS projects is not a straight forward task, at the moment of writing this paper, there is no research about this. Furthermore, there is little to no research about how women, or other minorities, feel about code of conduct, and how code of conduct impacts minorities in OSS projects. How a code of conduct is enforced can mostly be studied by talking to communities, to see how they perceive the enforcement, and to the maintainers enforcing it, which is something left for future work.

Another possible explanation, for the lack of impact that code of conduct has, in the long run, on proportion of women, and on time women remain contributors in OSS projects, might be due to the way women decide to contribute in a project. As concluded by Balali et al. [3], women decide to contribute in a project because they have a real interest in it or the community around it, or because women find the project "exciting" [3]. At the moment of writing this thesis, there is no research with respect to which communities adopt a code of conduct and whether they are of interest for women. It might be that projects analysed in this thesis are not of interest for women, or are not "exciting". It might be that projects which have had problems in the past, with offensive behaviour, are more likely to adopt a code of conduct, which again, might not seem interesting for women. Furthermore, there is little to no research about what women look for when deciding to contribute in a project and what women find "exciting" [3] in a project. Determining what women look for when deciding to contribute in a project, and what they find "exciting", can mostly be studied by talking to women, which is something left for future work. Knowing what projects women are interested in, might help in identifying new projects for analysing the impact of code of conduct.

The results obtained in this thesis, with respect to the lack of impact codes of conduct has on proportion of women in OSS projects, are the same as the ones obtained by Robson in his work [43], despite solving some of the threats found in Robson's work. As mentioned in Chapter 2.1, this thesis has taken into account more projects than in Robson's work, a different definition of contribution was also used, and the impact that time might have on results, was also taken into account. Confounds were also taken into account when doing the statistical analysis. Thus, despite removing some of the threats from Robson's work, the results remain the same as Robson's, indicating that code of conduct might indeed have no impact on proportion of women.

In terms of limitations of the results, only 418 projects were analyzed, out of 139,486 which have adopted a code of conduct. The low number of projects was due to filtering, such as removing non-engineered projects, clones, and projects which do not have a model 3 or 4 of contributorship. Adopting a different definition of contributorship, might have lead to different results since more than 50,000 projects were removed because they did not fulfill model 3 or 4 of contributorship. Also, because GitHub limits the search results to only 1000 answers, 17,514 projects could not be extracted from the platform. Because of the pandemic, 45,221 projects were not considered since they might influence the results. By far, the biggest number of projects removed, 69,888, were the ones which did not have a model 3 or 4 of contributorship. Since the projects removed were not taken into account, we cannot say anything about what impact they might have had on the proportion of women and time women remained contributors in a project. However, we believe that such filtering was needed to ensure the results obtained are not influenced by known external factors. Another possible limitation is using only GitHub projects in this thesis. We cannot say anything about what happens in projects from other platforms, such as GitLab and BitBucket.

To better understand the perception of women about code of conduct, and the impact that code of conduct has on women, interviews with women should be conducted. Conducting interviews with women might confirm if indeed, code of conduct does not have an impact on the proportion of women or on time women remain contributors in an OSS project. As mentioned throughout this section, talking with women might also help in understanding the reasons why they decide to contribute in a project and might help in including other projects in the analysis, or in removing some of the projects identified in Chapter 3 from the analysis, which might influence our results. Furthermore, the results from the interviews might help to take into account threats which were previously unknown. The results from interviews might also help in improving the statistical analysis and reduce confounds. The reason for conducting interviews with women is due to the lack of research on how women perceive code of conduct and what impact code of conduct has on women. At the time of writing this thesis, the paper written by Singh et al. [48], is the only qualitative research made about women's perception on code of conduct, but it only relies on analysing women's messages from online forums.

4.4 Threats to validity

The threats to validity for each of the research questions, have been discussed in Sections 4.1.4 and 4.2.2. In this section, a discussion about the overall threats of this thesis is provided, using the threats model proposed by Wohlin et al. [61].

Construct validity: the definition of contributorship, used in this thesis to identify contributors, might influence the results. In this thesis, it was decided to take into account the contribution model defined by Young et al. [62], and not restrict the contribution only to commits or pull requests, since, as suggested by Young et al. [62], commits and pull requests do not include all the contribution made in projects. Furthermore, to determine whether code of conduct has an impact on women's involvement, this thesis focuses on proportion of women, and time women remain contributors in a project, as a way to measure women's involvement. There might be other ways to measure women's involvement.

Internal validity: using quantitative method to determine the impact that code of conduct has on women's involvement. The results of this thesis could be further enhanced with a qualitative study, such as an interview, to understand how code of conduct influences them. The decision of using the contributorship models defined by Young et al. [62], which lead to a large proportion of projects being removed. These projects might have influenced the results obtained.

Maintainers of projects which have a code of conduct file, might not enforce the rules in their community, which might influence the effectiveness of code of conduct. As shown in the paper by Singh et al. [48], besides having a code of conduct in OSS projects, it is important how the code of conduct is enforced in OSS projects. To minimize this, each of the code of conduct files were manually analyzed to determine if a contact information is available for reporting violations of code of conduct. Whether reporting a violation is taking seriously or not (i.e. whether it is enforced), it is unknown. Given the lack of research in the field of codes of conduct, at the moment of writing this thesis, there is no known method in determining, from a quantitative way, whether a code of conduct is enforced or not.

The presence of confounds is another threat which might influence the results. The statistical models, used for answering **RQ1** and **RQ2**, take into account additional features. Only features based on which there is scientific proof that might influence the involvement of women, were taken into account. There might be other, unknown factors which might influence the results.

For identifying the gender of contributors, only their name and location were taken into account. Unfortunately, GitHub does not provide any method for users to select their gender in their GitHub profiles. The tool Namsor was used for identifying genders. At the moment of writing this thesis, there are no other tools for identifying the gender based on name, while also taking into account the country of origin and the different alphabets used in the world.

For each contributor who has a name, a gender is determined with a certain degree of confidence. The choice of the confidence level might represent another internal threat. To reduce any bias in selecting the confidence level, an interview with the creator of Namsor was done to determine what the best confidence level should be. As a result of this interview, a 60% confidence level was chosen. Furthermore, the entire statistical analysis was re-done using an 80% confidence level, and the same results and conclusions were drawn. Thus, the findings are stable.

External validity: only repositories from GitHub have been analyzed. GitHub is not the only online platform for hosting OSS projects, there are many others, such as GitLab and BitBucket. Furthermore, because of the extensive filtering done in Section 3.6, only a small number of repositories have been analyzed, making it difficult to generalize the results over the entire set. Also, only repositories which fulfill model 2, 3 and 4 of contributorship were kept, resulting in a small dataset. Removing non-engineered projects might also introduce bias, as shown in the paper by Munaiah et al. [31]. Thus, there is a large proportion of dataset that was not analyzed. This research started with 139,486 total projects, out of which only 418 projects were kept and analyzed. The results might not generalize to all projects which have adopted a code of conduct, but they generalize the software engineered projects which fulfill model 2, 3 and 4 of contributorship. To minimize removing projects which are suitable for analysis, all the decisions, such as where the code of conduct should be stored to be considered part of the project and not part of an external module, or excluding contributions after the pandemic started, were based on either previous researches, or on sampling the dataset and manually analyzing it to determine which is the best decision.

Chapter 5

Conclusion

In this thesis, a quantitative study is done to determine the impact that code of conduct has on women's involvement. At the moment of writing this thesis, there is little to no research about the impact that code of conduct has on women's involvement. The qualitative research done by Singh et al. [48], suggests that women engage more in projects which have adopted a code of conduct, because they feel that they belong in the community. On the contrary, the work done by Robson [43], concludes that code of conduct does not have any impact on proportion of women in OSS projects. To shed some light on whether code of conduct has an impact on women's involvement, we propose a quantitative analysis to determine whether code of conduct has an impact on women's involvement.

The focus of this thesis is on how code of conduct influences the proportion of women in OSS projects, as well as the time women remain contributors in a project. To provide an answer to the two aforementioned topics, GitHub projects which have adopted a code of conduct were identified using the GitHub search. From the identified projects, information belonging to model 2, and model 3 and 4 of contributorship, as defined by Young et al. [62], was extracted from GitHub, in order to determine the proportion of women in OSS projects as well as the time women remain contributors in a project. The gender of the contributors was determined based on the names of the contributors because GitHub does not provide any information on their platform about the gender of their users. To infer the gender from contributors' names, Namsor was used. To increase the accuracy of Namsor, the location, that contributors have made available on their profile, was used as well. Because of the pandemic, all GitHub data generated after 01-March-2020 (i.e. the date when the pandemic started), was discarded from our analysis.

To determine whether code of conduct has an influence on the proportion of women in OSS projects, an RDD technique was applied, on the data belonging to model 2 of contributorship, in order to take into account the passing of time. Because of the few data for model 3 and 4 of contributorship, a linear regression technique was applied. The results concluded that in the long run, there is no statistical significant evidence that code of conduct has an impact on the percentage of women in OSS projects. In the month that the code of conduct was adopted, the percentage of women has increased by 0.01627%.

Besides analyzing the proportion of women, the time women remain contributors in a project was also analyzed using a survival analysis, for plotting the survival curves. In addition to the survival analysis, a log rank test was used to determine if the survival curves are different, as well as a Cox-proportional statistical model, to take into account for confounds. The conclusion was that code of conduct does not have any impact on the time women remain contributors in a project.

The findings suggest that there is no statistical significant evidence that code of conduct has an impact on women's involvement. Further research is needed to confirm if this is the case or not, as well as an extension of the research, to include other measurements for involvement, besides percentage of women and time women remain contributors in a project.

5.1 Future work

In this section possible future work is discussed. In Section 5.1.1, future work which might improve the work done in this thesis, is discussed, while in Section 5.1.2, future work for expanding the research, beyond the focus of this thesis, is discussed.

5.1.1 Improve current work

The results obtained in this thesis should be augmented with a qualitative study to determine how the code of conduct impacts women's involvement. Interviews with women, who contribute in OSS projects, should be conducted to confirm whether code of conduct does not have any impact on the proportion of women who contribute, and no impact on the time women remain contributors. Based on possible insights obtained from the interviews, the study can be adjusted accordingly.

Another point to improve the current work is related with the way involvement is measured. In this thesis, the focus was on proportion of women in OSS projects and on time women remain contributors in OSS projects. There might be other ways to measure involvement, which might lead to different results. Furthermore, it might be that women are influenced in other ways, unrelated to contribution, by code of conduct. To determine what measurements should be used to determine whether code of conduct has any influence on women who contribute in OSS projects, interviews with women should be conducted.

Data was extracted only from GitHub. There might be projects in which people can make contributions outside of GitHub platform. For example, projects might use Slack, Discord, emails or other platforms in which they can engage with the OSS project's community and contribute towards the project. Contribution, for GitHub projects, made on other platforms is not taken into account in this thesis. Future work should take this into account, since not all contribution can be seen on GitHub, as shown in the paper by Young et al. [62].

In the work by Singh et al. [48], women reported that whether code of conduct is enforced or not, in the OSS project, has an impact on how they engage in the community. In this thesis, one factor was taken into account which might reflect whether code of conduct is enforced or not in an OSS project: presence of contact information to report violations of code of conduct. At the time of writing this thesis, there is no research that we are aware of, which studies what are the signs of a code of conduct being enforced in the community. Performing such a study might be useful in differentiating projects in which the code of conduct is enforced, from the ones in which code of conduct is not enforced. Since there are codes of conduct written by different individuals and organizations, it might be that women have different perceptions about each one of the codes of conduct, or it might be that some codes of conduct, which contain certain information, might have a bigger impact on women in comparison to other codes of conduct. To determine whether women view certain codes of conduct, differently than others, interviews should be conducted with women.

5.1.2 Extend the area of research

Women are not the only minority in OSS projects. Code of conduct is aimed towards all minorities, not only women. A possible extension of this study would be to include other minorities and determine how code of conduct has influenced them. It might be that code of conduct has more impact on some minorities, while no impact on other minorities.

In this thesis, only projects stored on GitHub were considered. Other platforms, such as GitLab, can be included in the thesis, for better generalizing the results.

Bibliography

[1] Alison Adam, Debra Howcroft, and Helen Richardson. A decade of neglect: reflecting on gender and is. *New Technology, Work and Employment*, 19(3):222–240, 2004.

[2] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346, 2003.

[3] Sogol Balali, Igor Steinmacher, Umayal Annamalai, Anita Sarma, and Marco Aurelio Gerosa. Newcomers' barriers... is that all? an analysis of mentors' and newcomers' barriers in oss projects. *Computer Supported Cooperative Work (CSCW)*, 27(3):679–714, 2018.

[4] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[5] Marcus Vinicius Bertoncello, Gustavo Pinto, Igor Scaliante Wiese, and Igor Steinmacher. Pull requests or commits? which method should we use to study contributors' behavior? In *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 592–601. IEEE, 2020.

[6] Kelly Blincoe, Olga Springer, and Michal Wrobel. Perceptions of gender diversity's impact on mood in software development teams. *Ieee Software*, 36(5):51–56, 2019.

[7] Hudson Borges, Marco Tulio Valente, Andre Hora, and Jailton Coelho. On the popularity of github applications: A preliminary note. *arXiv preprint arXiv:1507.00604*, 2015.

[8] Kay Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, Steven L Scott, et al. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9(1):247–274, 2015.

[9] Donald Campbell and Thomas Cook. Quasi-experimentation. *Chicago, IL: Rand Mc-Nally*, 1979.

[10] Edna Dias Canedo, Rodrigo Bonifácio, Márcio Vinicius Okimoto, Alexander Serebrenik, Gustavo Pinto, and Eduardo Monteiro. Work practices and perceptions from women core developers in oss communities. In *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–11, 2020.

[11] Edna Dias Canedo, Heloise Acco Tives, Madianita Bogo Marioti, Fabiano Fagundes, and José Antonio Siqueira de Cerqueira. Barriers faced by women in software development projects. *Information*, 10(10):309, 2019.

[12] Sherae Daniel, Ritu Agarwal, and Katherine Stewart. The effects of diversity in global, distributed collectives: A study of open source project success. *Information Systems Research*, 24(2):312–333, 2013.

[13] Tapajit Dey, Sara Mousavi, Eduardo Ponce, Tanner Fry, Bogdan Vasilescu, Anna Filippova, and Audris Mockus. Detecting and characterizing bots that commit code. In *Proceedings of the 17th international conference on mining software repositories*, pages 209–219, 2020.

[14] Patrick Erwin. Corporate codes of conduct: The effects of code content and quality on ethical performance. *Journal of Business Ethics*, 99(4):535–548, 2011.

[15] Davide Falessi, Wyatt Smith, and Alexander Serebrenik. Stress: A semi-automated, fully replicable approach for project selection. In *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 151–156. IEEE, 2017.

[16] Denae Ford, Alisse Harkins, and Chris Parnin. Someone like me: How does peer parity influence participation of women on stack overflow? In *2017 IEEE symposium on visual languages and human-centric computing (VL/HCC)*, pages 239–243. IEEE, 2017.

[17] Rishab Ghosh, Ruediger Glott, Bernhard Krieger, and Gregorio Robles. Free/libre and open source software: Survey and study. 2002.

[18] Mehdi Golzadeh, Alexandre Decan, Damien Legay, and Tom Mens. A ground-truth dataset and classification model for detecting bots in github issue and pr comments. *Journal of Systems and Software*, 175:110911, 2021.

[19] Georgios Gousios. The ghtorent dataset and tool suite. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, pages 233–236. IEEE, 2013.

[20] Susan C Herring and John C Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2006.

[21] Michael Hilton, Timothy Tunnell, Kai Huang, Darko Marinov, and Danny Dig. Usage, costs, and benefits of continuous integration in open-source projects. In *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 426–437. IEEE, 2016.

[22] Guido Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

[23] Nasif Imtiaz, Justin Middleton, Joymallya Chakraborty, Neill Robson, Gina Bai, and Emerson Murphy-Hill. Investigating the effects of gender bias on github. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 700–711. IEEE, 2019.

[24] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel German, and Daniela Damian. The promises and perils of mining github. In *Proceedings of the 11th working conference on mining software repositories*, pages 92–101, 2014.

[25] Christiana Kartsonaki. Survival analysis. *Diagnostic Histopathology*, 22(7):263–270, 2016.

[26] Stefan Krüger and Ben Hermann. Can an online service predict gender? on the state-of-the-art in gender identification from texts. In *2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE)*, pages 13–16. IEEE, 2019.

[27] Amanda Lee and Jeffrey Carver. Floss participants' perceptions about gender and inclusiveness: a survey. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 677–687. IEEE, 2019.

[28] Xiaoguang Lu, Hong Chen, and Anil K Jain. Multimodal facial gender and ethnicity identification. In *International conference on biometrics*, pages 554–561. Springer, 2006.

[29] Jonathan Marshall. Online life and gender vagueness and impersonation. In *Encyclopedia of gender and information technology*, pages 932–938. IGI Global, 2006.

[30] Warwick McKibbin, Roshen Fernando, et al. The economic impact of covid-19. *Economics in the Time of COVID-19*, 45(10.1162), 2020.

[31] Nuthan Munaiah, Steven Kroh, Craig Cabrey, and Meiyappan Nagappan. Curating github for engineered software projects. *Empirical Software Engineering*, 22(6):3219–3253, 2017.

[32] D Nafus, James Leach, and B Krieger. Flosspols deliverable d 16 gender: Integrated report of findings. 01 2006.

[33] Dawn Nafus. 'patches don't have gender': What is not open in open source software. *New Media & Society*, 14(4):669–683, 2012.

[34] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2):133–142, 2013.

[35] Paulo Anselmo da Mota Silveira Neto, Umme Ayda Mannan, Eduardo Santana de Almeida, Nachiappan Nagappan, David Lo, Pavneet Singh Kochhar, Cuiyun Gao, and Iftekhar Ahmed. A deep dive on the impact of covid-19 in software development. *arXiv preprint arXiv:2008.07048*, 2020.

[36] Edson Oliveira, Eduardo Fernandes, Igor Steinmacher, Marco Cristo, Tayana Conte, and Alessandro Garcia. Code and commit metrics of developer productivity: a study on team leaders perceptions. *Empirical Software Engineering*, 25(4):2519–2549, 2020.

[37] Mohd Hafeez Osman, Truong Ho-Quang, and Michel Chaudron. An automated approach for classifying reverse-engineered and forward-engineered uml class diagrams. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 396–399. IEEE, 2018.

[38] Peterson Ozili and Thankom Arun. Spillover of covid-19: impact on the global economy. *Available at SSRN 3562570*, 2020.

[39] Dewayne Perry, Adam Porter, and Lawrence Votta. Empirical studies of software engineering: a roadmap. In *Proceedings of the conference on The future of Software engineering*, pages 345–355, 2000.

[40] Antoine Pietri, Guillaume Rousseau, and Stefano Zacchiroli. Forking without clicking: on how to identify software repository forks. In *Proceedings of the 17th International Conference on Mining Software Repositories*, pages 277–287, 2020.

[41] Huilian Sophie Qiu, Yucen Lily Li, Susmita Padala, Anita Sarma, and Bogdan Vasilescu. The signals that potential contributors look for when choosing open-source projects. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–29, 2019.

[42] Huilian Sophie Qiu, Alexander Nolte, Anita Brown, Alexander Serebrenik, and Bogdan Vasilescu. Going farther together: The impact of social capital on sustained participation in open source. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 688–699. IEEE, 2019.

[43] Neill Robson. Diversity and decorum in open source communities. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 986–987, 2018.

[44] Nigel Ross. Writing in the information age. *English Today*, 22(3):39–45, 2006.

[45] Simon Sheather. *A modern approach to regression with R*. Springer Science & Business Media, 2009.

[46] Marcelino Campos Oliveira Silva, Marco Tulio Valente, and Ricardo Terra. Does technical debt lead to the rejection of pull requests? *arXiv preprint arXiv:1604.01450*, 2016.

[47] Jaspreet Singh and Jagandeep Singh. Covid-19 and its impact on society. *Electronic Research Journal of Social Sciences and Humanities*, 2, 2020.

[48] Vandana Singh, Brice Bongiovanni, and William Brandon. Codes of conduct in open source software—for warm and fuzzy feelings or equality in community? *Software Quality Journal*, pages 1–40, 2021.

[49] Klaas-Jan Stol and Brian Fitzgerald. The abc of software engineering research. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 27(3):1–51, 2018.

[50] John R Suler. Do boys and girls just wanna have fun. *Gender communication*, pages 149–153, 2004.

[51] Donald Thistlethwaite and Donald Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.

[52] Parastou Tourani, Bram Adams, and Alexander Serebrenik. Code of conduct in open source projects. In *2017 IEEE 24th international conference on software analysis, evolution and reengineering (SANER)*, pages 24–33. IEEE, 2017.

[53] Asher Trockman, Shurui Zhou, Christian Kästner, and Bogdan Vasilescu. Adding sparkle to social coding: an empirical study of repository badges in the npm ecosystem. In *Proceedings of the 40th International Conference on Software Engineering*, pages 511–522, 2018.

[54] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*, pages 332–338. IEEE, 2012.

[55] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. Perceptions of diversity on git hub: A user survey. In *2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering*, pages 50–56. IEEE, 2015.

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

[57] Bogdan Vasilescu, Alexander Serebrenik, and Vladimir Filkov. A data set for social diversity studies of github teams. In *2015 IEEE/ACM 12th working conference on mining software repositories*, pages 514–517. IEEE, 2015.

[58] Bogdan Vasilescu, Stef Van Schuylenburg, Jules Wulms, Alexander Serebrenik, and Mark van den Brand. Continuous integration in a social-coding world: Empirical evidence from github. In *2014 IEEE international conference on software maintenance and evolution*, pages 401–405. IEEE, 2014.

[59] Yi Wang and David Redmiles. Implicit gender biases in professional software development: An empirical study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 1–10. IEEE, 2019.

[60] Shi-Hong Weng, Anna Ya Ni, Alfred Tat-Kei Ho, and Ruo-Xi Zhong. Responding to the coronavirus pandemic: a tale of two cities. *The American Review of Public Administration*, 50(6-7):497–504, 2020.

[61] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[62] Jean-Gabriel Young, Amanda Casari, Katie McLaughlin, Milo Z Trujillo, Laurent Hébert-Dufresne, and James P Bagrow. Which contributions count? analysis of attribution in open source. *arXiv preprint arXiv:2103.11007*, 2021.

[63] Yangyang Zhao, Alexander Serebrenik, Yuming Zhou, Vladimir Filkov, and Bogdan Vasilescu. The impact of continuous integration on other software development practices: a large-scale empirical study. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 60–71. IEEE, 2017.

[64] Jiaxin Zhu, Minghui Zhou, and Audris Mockus. Patterns of folder use and project popularity: A case study of github repositories. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 1–4, 2014.

Appendix A

GitHub searches

In this section, the results, obtained when identifying repositories which have a code of conduct are presented.

A.1 The most common 30 file names returned by GitHub when searched by code for “code of conduct”

The GitHub search by code: *“code of conduct”*

Overall stats:
Results GitHub found: 1,378,825
Results retrieved: 700,184 (50.78%)
Unique projects retrieved: 330,608
Results missed because of GitHub limitations: 678,274 (49.19%)

README.md: 242,492 (34.63%)
CODE_OF_CONDUCT.md: 168,840 (24.11%)
CONTRIBUTING.md: 148,118 (21.15%)
CHANGELOG.md: 26,321 (3.75%)
CONDUCT.md: 10,007 (1.42%)
code-of-conduct.md: 9,772 (1.39%)
readme.md: 8,488 (1.21%)
contributing.md: 5,729 (0.81%)
History.md: 4,961
CODE-OF-CONDUCT.md: 4,402
code_of_conduct.md: 4,381
index.md: 4,277
README.es.md: 4,195
Readme.md: 3,412
ISSUE_TEMPLATE.md: 2,608
PULL_REQUEST_TEMPLATE.md: 2,554
Changelog.md: 1,678
links.md: 1,647
Code-of-Conduct.md: 1,624
bug_report.md: 1,095
feature_request.md: 1,053
issue_template.md: 926
conduct.md: 850

coc.md:	716
Contributing.md:	630
purplebooth.md:	438
COLLABORATOR_GUIDE.md:	428
support.md:	418
helpers.md:	394
changelog.md:	391

A.2 Number of path depths for each of the code of conduct files found in GitHub search

The GitHub search by code:
filename:"code of conduct" extension:md
filename:"conduct" extension:md size:XX..YY

Overall stats:
Results GitHub found: 704,965
Results retrieved: 232,702 (33.01%)
Unique projects retrieved: 177,461
Results missed because of GitHub limitations: 471,965 (66.94%)

GitHub results with path length 0: 85,530. Sample size: 96.
GitHub results with path length 1: 19,128. Sample size: 96.
GitHub results with path length 2: 35,191. Sample size: 96.
GitHub results with path length 3: 38,377. Sample size: 96.
GitHub results with path length 4: 26,775. Sample size: 96.
GitHub results with path length 5: 10,059. Sample size: 95.
GitHub results with path length 6: 10,320. Sample size: 95.
GitHub results with path length 7: 2,827. Sample size: 93.
GitHub results with path length 8: 1,177. Sample size: 89.
GitHub results with path length 10: 815. Sample size: 86.
GitHub results with path length 11: 152. Have analyzed entire set.
GitHub results with path length 12: 82. Have analyzed entire set.
GitHub results with path length 13: 42. Have analyzed entire set.
GitHub results with path length 14: 131. Have analyzed entire set.
GitHub results with path length 15: 77. Have analyzed entire set.
GitHub results with path length 16: 18. Have analyzed entire set.
GitHub results with path length 17: 2. Have analyzed entire set.
GitHub results with path length 18: 12. Have analyzed entire set.
GitHub results with path length 19: 1. Have analyzed entire set.
GitHub results with path length 20: 3. Have analyzed entire set.
GitHub results with path length 21: 1. Have analyzed entire set.
GitHub results with path length 22: 1. Have analyzed entire set.

To determine whether a code of conduct belongs to the project, and not to external libraries, a path threshold was set. For determining the path threshold, each of the path lengths was manually analyzed by taking a sample with confidence level 95% and confidence interval 10%. All code of conducts in paths with length higher than threshold, are considered to be part of external libraries. Code of conducts which are part of external libraries are removed from the dataset.

A.3 The most common paths of depth zero or one for each of the code of conduct files found in GitHub search

The GitHub search by code:
filename:"code of conduct" extension:md
filename:"conduct" extension:md

NOTE: the dataset used is the same as the one in appendix A.2, just that all the results which have path length higher than one were removed.

Out of 232,702 results, only 104,658 results had a path length smaller or equal with one. Thus, 128,043 results were removed. In terms of unique projects: 102,358 out of a total of 177,461 unique projects had path depth smaller or equal to one. Thus, 75,076 unique projects were removed.

List with the most common paths of depth zero or one:
CODE_OF_CONDUCT.md: 77402 (73.59%)
.github/CODE_OF_CONDUCT.md: 8286 (7.91%)
code-of-conduct.md: 3722
conduct.md: 1798
docs/CODE_OF_CONDUCT.md: 1680
code_of_conduct.md: 1578
.github/conduct.md: 647
docs/code-of-conduct.md: 463
docs/conduct.md: 238
trellis/CODE_OF_CONDUCT.md: 160
site/CODE_OF_CONDUCT.md: 159
Code-of-Conduct.md: 143
Code_of_Conduct.md: 130
.oh-my-zsh/CODE_OF_CONDUCT.md: 112
.github/CODE-OF-CONDUCT.md: 100
CodeOfConduct.md: 98
sendgrid-php/CODE_OF_CONDUCT.md: 95
.github/code-of-conduct.md: 80
bootstrap/CODE_OF_CONDUCT.md: 79
app/CODE_OF_CONDUCT.md: 71
src/CODE_OF_CONDUCT.md: 67
Code of Conduct.md: 66
devtools/CODE_OF_CONDUCT.md: 65
content/code-of-conduct.md: 60
fabric-samples/CODE_OF_CONDUCT.md: 59
update/CODE_OF_CONDUCT.md: 58
_site/CODE_OF_CONDUCT.md: 57
CODE_OF_CONDUCT_ES.md: 52
other/CODE_OF_CONDUCT.md: 49
bootstrap-4.5.3/CODE_OF_CONDUCT.md: 46

Summarizing the path stats, including the percentages out of all files retrieved in appendix A.2 with path length smaller or equal with one, sorted based on results:
Codes of conduct stored in the root path: 84,989 (81.2%)
Codes of conduct stored in the .github folder: 9,113 (8.7%)
Codes of conduct stored in the docs folder: 2,381 (2.27%)
Codes of conduct stored in the trellis folder: 160 (0.15%)
Codes of conduct stored in other folders: 7,855 (7.5%)

Appendix B

Data used for statistics

In this section, information about the dataset used for statistics is presented.

B.1 RQ 1: Overall statistics for projects which have a model 2 contribution made by women in the pre-conduct or post-conduct period

The projects obtained at the end of chapter 3, which do not have any women's involvement in any of the pre-conduct or post-conduct periods, are removed since there is nothing to be measured for these projects. Information about the projects which were kept in the dataset can be found below.

Number of repos with females: 294
Unique logins (male / female / no gender / no name): 114,072
Unique logins no name: 16,172 (from the 23,965 defined in step_18)
Unique logins no gender: 17,191 (from the 25,224 defined in step_18)
Unique logins females: 6,985 (from the 10,230 defined in step_18)
Unique logins males: 72,986 (from the 106,812 defined in step_18)
Out of all logins (114,072), 85.82% have a name (97,900)
Out of all logins which have a name (97,900), 81.68% have a gender defined (79,971)
Out of all logins which have a gender (79,971), 8.73% are females (6,985)
All model 2 contributions (male / female / not defined / no name): 6,885,630
Number of total model 2 contributions made by no name: 116,342
Number of total model 2 contributions made by no gender: 1,120,101
Number of total model 2 contributions made by females: 592,494
Number of total model 2 contributions made by males: 4,963,607
Out of all model 2 contributions (6,885,630), 98.31% have a name (6,769,288)
Out of all model 2 contributions which have a name (6,769,288), 82.07% have a gender defined (5,556,101)
Out of all model 2 contributions which have a gender (5,556,101), 10.62% are females (592,494)

B.2 RQ 1: Overall statistics for projects which have a model 2 contribution made by women in the defined period length and with period length larger or equal to 360 days

The projects, from B.1, which have either a time length period smaller than 360 days, or no contribution made by women in any of the 12 time points before or after the code of conduct was adopted, are removed from the dataset. Information about the projects which were kept in the dataset can be found below.

Number of repos with females: 149
Unique logins (male / female / no gender / no name): 67,938
Unique logins no name: 7,988
Unique logins no gender: 10,378
Unique logins females: 4,124
Unique logins males: 44,992
Out of all logins (67,938), 88.25% have a name (59,950)
Out of all logins which have a name (59,950), 82.68% have a gender defined (49,572)
Out of all logins which have a gender (49,572), 8.31% are females (4,124)
All model 2 contributions (male / female / not defined / no name): 4,128,537
Number of total model 2 contributions made by no name: 55,934
Number of total model 2 contributions made by no gender: 684,028
Number of total model 2 contributions made by females: 384,064
Number of total model 2 contributions made by males: 2,980,324
Out of all model 2 contributions (4,128,537), 98.6% have a name (4,072,603)
Out of all model 2 contributions which have a name (4,072,603), 83.2% have a gender defined (3,388,575)
Out of all model 2 contributions which have a gender (3,388,575), 11.33% are females (384,064)