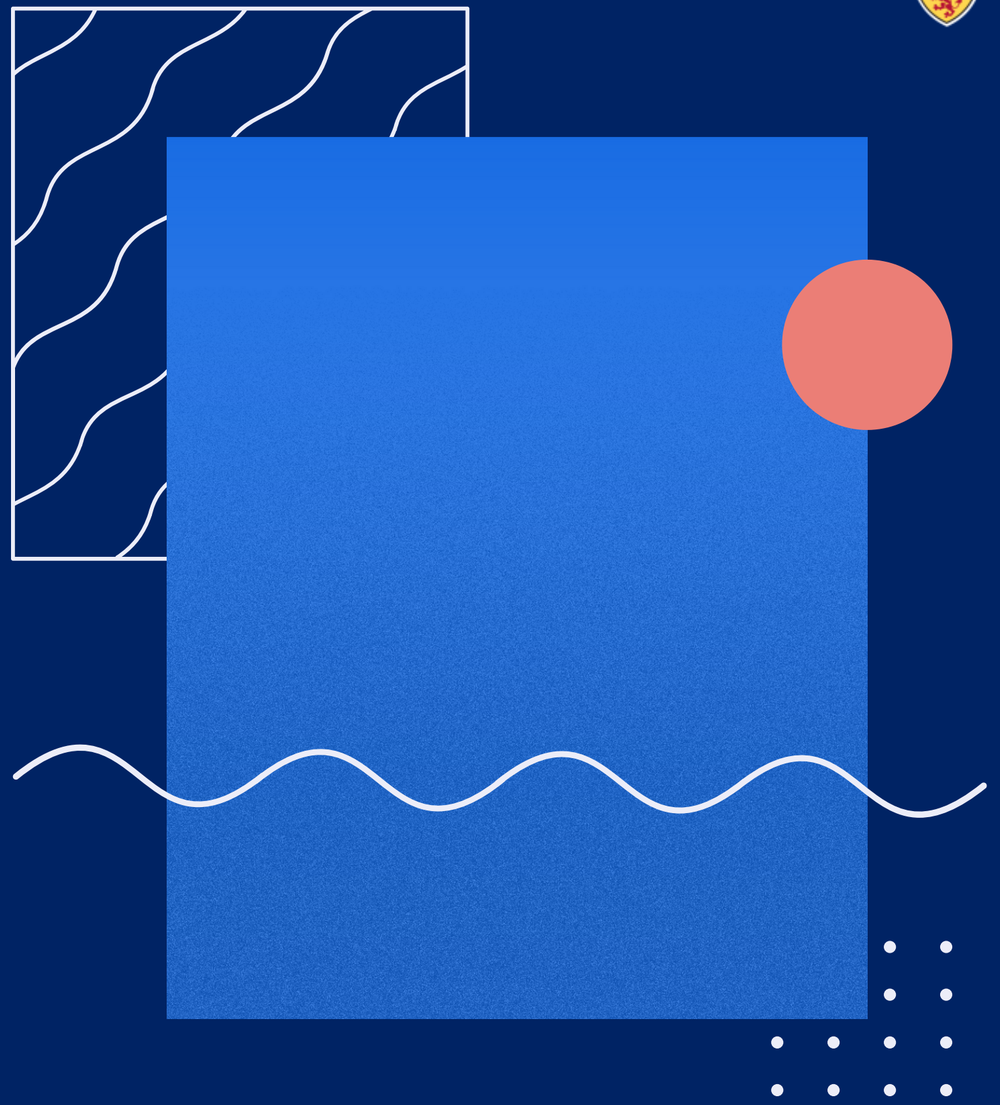
AN EMPIRICAL STUDY OF FIRST TIME OPEN SOURCE CONTRIBUTORS ON GITHUB

VIKRAM N. SUBRAMANIAN
SWAG LAB, UNIVERSITY OF WATERLOO



"finding a task to start with was the most common difficulty faced by first time contributors"

The Hard Life of Open Source Software Project Newcomers

Igor Steinmacher,
Igor Scaliante Wiese
DACOM – UTFPR
Campo Mourão-PR - Brazil
{igorfs, igor}@utfpr.edu.br

Tayana Conte
IComp – UFAM
Manaus-AM - Brazil
tayana@icomp.ufam.edu.br

Marco Aurelio Gerosa
IME – USP
São Paulo-SP - Brazil
gerosa@ime.usp.br

David Redmiles
Department of Informatics
Univ. of California, Irvine
Irvine-CA - USA
redmiles@ics.uci.edu

ABSTRACT

While onboarding an open source software (OSS) project, contributors face many different barriers that hinder their contribution, leading in many cases to dropouts. Many projects leverage the contribution of outsiders and the sustainability of the project relies on retaining some of these newcomers. In this paper, we discuss some barriers faced by newcomers to OSS. The barriers were identified using a qualitative analysis on data obtained from newcomers and members of OSS projects. We organize the results in a conceptual model composed of 38 barriers, grouped into seven different categories. These barriers may motivate new studies and the development of appropriate tooling to better understand and support the onboarding of new contributors.

developer wants to become a contributor, committer, or a core member, although all of them are subject to the problems of onboarding before making their first contribution.

Dagenais et al. [9], for example, compare software project newcomers to explorers who need to orient themselves in a hostile environment. On the one hand, newcomers need to learn social and technical aspects alone, exploiting existing information in mailing lists, source code repositories, and issue managers [29]. On the other hand, it is not easy to access this information due to the large volume, lack of tools to navigate the repository, and the difficulty of linking logically related items in different sources [8].

OSS projects can benefit from more contributions if they offer the right support to newcomers during their onboarding. To achieve this, it is necessary to understand how OSS communities interact

Motivation

To better understand contributions by first time contributors by studying the characteristics of the first pull request (PR) made to an OSS project by them.

Data Collection

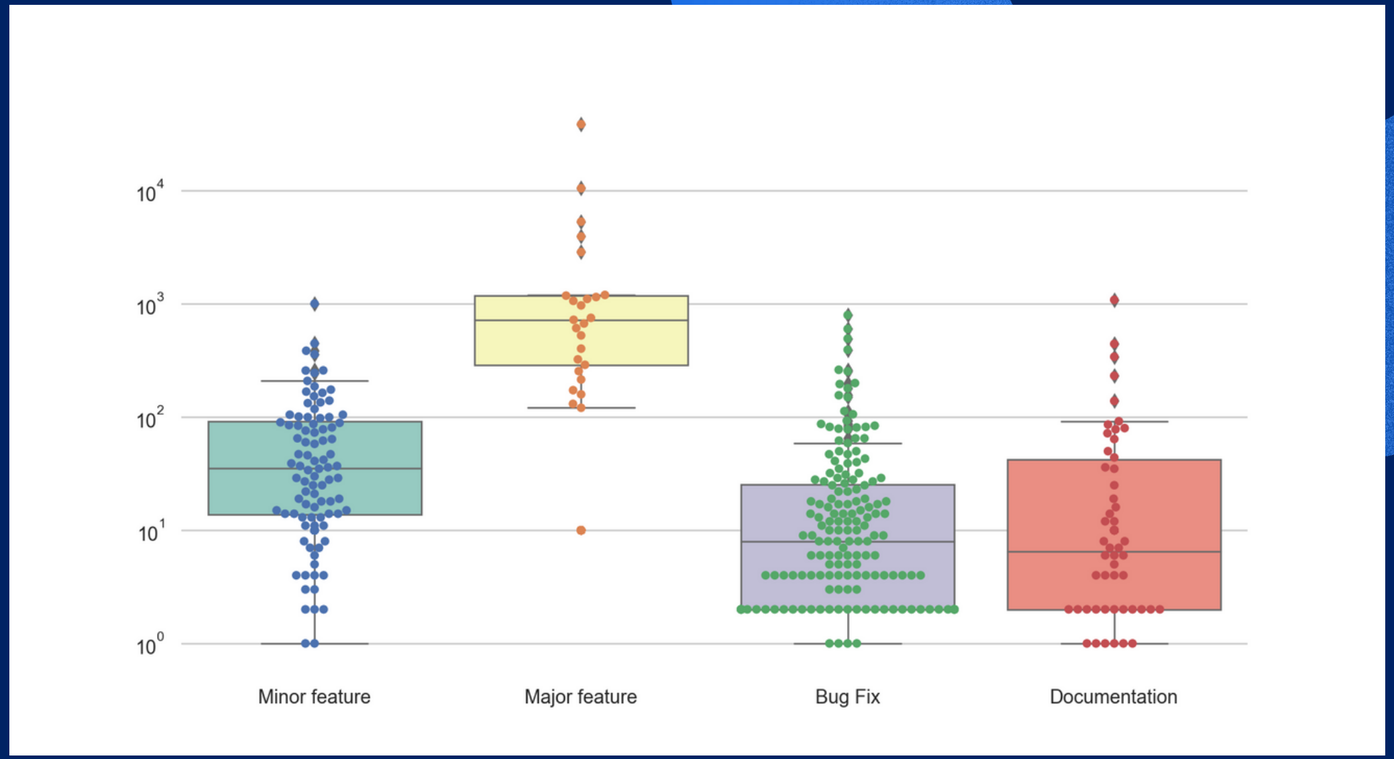
We use the GitHub API and our own logic to obtain the first PR of 3501 users.
Briefly-

1. We first get a list of 1000 repositories from The Apache Software Foundation's GitHub page.
2. We then get the top 1000 contributors to these repositories.
3. We then study the repositories these users have forked (and their parents) to identify any open source contributions and eventually their first.

Results

31.5% of the changes were 1-5 lines and 18.8% of the changes were 6-15 lines with some extreme outliers.

54.77% of all files changed were modified/edited, 35.41% were added, 5.85% were renamed and 3.94% were removed



UNIVERSITY OF WATERLOO

Results

FILE TYPE USED	
Java	34.67%
Python	5.93%
PNG- Picture Files	5.92%
Markdown	4.57%
JavaScript	4.26%
XML	4.12%
Go	3.56%
Shell	2.21%
Scala	2.13%

Java is by far the most popular language at 34.67%. This observation aligns with the fact that Java was the most popular language throughout the 2010s

Non-code contributions such as adding picture files were surprisingly popular

C++ is extremely unpopular and was only the 15th most popular language for first time contributions while it was the 4th most popular 'main' language for the OSS repositories we studied.

Takeaways

Most first timers take up small to medium sized tasks. However, they should not be discouraged to take up big tasks as the entire 4th quartile of changes were over 75 lines.

First time contributors should also look into non-code contributions.

Moderators of OSS projects could use our results to prioritize tasks based on the language of the file that needs to be modified for first time contributors.

Further work

	Count	Percentage of all changes	Lines changed - Mean	Lines changed - Median	Files changed - Mean	Files changed - Median
Documentation	53	13.25%	60.30	7	2.19	1
Feature	140	35.00%	839.57	52.5	6.24	2
Bugs	188	47.00%	37.85	8	2.13	1
Refactoring	2	0.50%	2051.00	2051	19.00	19
GIT related issues	2	0.50%	3.00	3	1.50	1.5
Test Cases	13	3.25%	244.46	21	3.92	2
Other	2	0.50 %	36.00	36	69.50	69.5

	Count	Percentage of all changes	Percentage of all bug fixes	Lines changed - Mean	Lines changed - Median	Files changed - Mean	Files changed - Median	Example
Typographic Bug	54	13.50%	28.72%	21.27	2	1.52	1	bit.ly/3ejFsxN
Exception Handling	6	1.50%	3.19%	6.83	4	1.17	1	bit.ly/2XxOn8X
Processing Error	2	0.50%	1.06%	3.50	3.5	1.00	1	bit.ly/2yhuGaH
Wrong Control Flow	8	2.00%	4.26%	50.25	16	2.75	2.5	bit.ly/3a6eB4X
Corner Cases	6	1.50%	3.19%	48.83	7.5	1.83	1.5	bit.ly/2xr11fj
Missing Cases	32	8.00%	17.02%	32.625	7.5	1.375	1	bit.ly/2V3rF70

Find out more: bit.ly/2Z1XpvJ