ECS 260 Project Proposal

Professor: Vladimir Filkov

Working Group: Bobby Missirian

Karamjeet Singh Gulati

January 24, 2024

1 Introduction

In an increasingly interconnected world, Open Source Software (OSS) has become the backbone of digital infrastructure, powering everything from individual applications to global systems. The Apache Software Foundation Incubator (ASFI) provides a structured environment to foster the growth and sustainability of OSS projects. ASFI helps OSS projects develop into sustainable communities so that the codebase is maintained into the forseeable future.

In open source projects, contributors often utilize forks to develop experimental features without interfering with the main branch, and then they try to merge their completed feature back into main. Sometimes, after significant effort on the part of the contributor, the admins reject their changes, leading to inefficiency [1].

2 Research Question

In 2021, Zhou et al. analyzed data from commits and pull requests to determine which forks are integrated into the main branch and which are abandoned [1]. They determined that centralization and modular structure in OSS communities are associated with more efficient forking practices. Given the relationship between efficient forking and structure, forking practices may also have implications for the sustainability of the project as a whole.

This research project examines how the way in which contributors use forking in OSS projects plays a pivotal role in the projects' trajectories. The questions guiding this research are:

1. What fork practices are correlated with the sustainability of OSS projects in ASFI?

This question aims to explore the correlation between specific forking practices, as documented by Zhou et al. (2019), and the sustainability of open-source software (OSS) projects within the Apache Software Foundation Incubator (ASFI), drawing on the sustainability metrics and insights provided by Yin et al. (2021) [1][2].

2. To what extent can forking practice factors predict the sustainability of ASFI projects?

This question seeks to quantify the predictive power of forking practices on the sustainability of ASFI projects. It builds on the findings of Zhou et al. (2019) and Yin et al. (2021) to examine whether the characteristics of forking practices can serve as reliable indicators for the future success and longevity of OSS projects in the ASFI[1][3].

3. Does forking lead to changes in development practices, commit frequency, or other process metrics?

Focused on the operational impacts of forking, this question investigates how forking practices influence various process metrics within OSS projects, including development practices and commit frequency. This inquiry is rooted in the broader context of OSS project management and sustainability, as highlighted in the studies by Zhou et al. (2019) and Xiao et al. (2023)[1][4].

3 Approach

3.1 Data

We will start from an ASFI Dataset of successful and unsuccessful projects. The dataset encompasses developer coding and communication activities from 269 projects, each given a status by AFSI as either "graduated" (successful) or "retired" (unsuccessful). The dataset includes detailed records of commits, email communications, file changes, and other collaborative activities within the ASFI projects, providing a comprehensive view of each project's lifecycle within ASFI [3]. We will track evidence of forks throughout the data to build a comprehensive picture of forking practice in each project. We will use the status of each project as an indicator of the sustainability of its development.

3.2 Methodology

Data Visualization: Look over the raw data relevant to our RQs, including aliases, commits, file list, project list, messages, people.

Preliminary Data Analysis: Apply the fork tracing algorithm that was run on selected Github projects in Zhou et al. to the ASFI dataset.

Exploratory Data Analysis: Exploratory data analysis: Conduct an original analysis of the data to test hypotheses relating to RQ 1.

Model Building and Testing: Model building and testing: Develop a model based on the most strongly correlated fork features. Then validate to insure accuracy and reliability. Use the model to test hypotheses relating to RQ 2.

Supplementary Analysis: Perform any further tests arising from RQ 3 and from observations in the previous phases, time permitting.

Results Compilation: Analyze the outputs of the models, draw conclusions as to all hypotheses, and compile the findings.

Paper Writing: Document the research process and findings in our paper.

4 Timeline

The project will adhere to a structured timeline, ensuring each phase is given the due diligence it deserves. Key milestones include:

Phase	Activity	Start Week	End Week
Preparation Phase	Data Visualization	1	2
	Preliminary Data Analysis	3	4
Analysis Phase	Exploratory Data Analysis	4	5
	Model Building and Testing	6	7
	Supplementary Analysis	7	8
Finalization Phase	Results Compilation	8	9
	Paper Writing	9	10
Milestone	Final Submission	10	10

Table 1: Schedule

5 Team Membership and Attestation

This proposal was made by Robert Missirian and Karamjeet Gulati. We are the only remaining students in our group.

References

[1] Shurui Zhou, Bogdan Vasilescu, and Christian Kästner. What the fork: A study of inefficient and efficient forking practices in social coding. In *Proceedings of the 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '19)*, page 12, New York, NY, USA, 2019. ACM.

[2] Likang Yin, Zhuangzhi Chen, Qi Xuan, and Vladimir Filkov. Sustainability forecasting for apache incubator projects. In *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21)*, page 12, New York, NY, USA, 2021. ACM.

[3] Likang Yin, Zhiyuan Zhang, Qi Xuan, and Vladimir Filkov. Apache software foundation incubator project sustainability dataset, 2021. Data provided by the Apache Software Foundation Incubator.

[4] Wenxin Xiao, Hao He, Weiwei Xu, Yuxia Zhang, and Minghui Zhou. How early participation determines long-term sustained activity in github projects?, 2023.