ECS 260 Project Progress Report

Professor: Vladimir Filkov

Working Group: Bobby Missirian

Karamjeet Singh Gulati

Pu Sun

February 27, 2024

1 Introduction

Our study explores the key factors influencing the success of OSS projects during their incubation at the ASFI, focusing on early activities, contribution diversity, and communication patterns. We aim to understand how these elements impact sustainability and success, offering insights for enhancing OSS project strategies. The theoretical foundation integrates concepts from OSS sustainability, the role of community dynamics, and structural contingency theory, emphasizing the nuanced interplay between project organization and socio-technical interactions.

1.1 Theory

We base our study on structural contingency theory. This theory says that an organization's structure, like OSS projects in the ASFI, depends on its internal and external contexts. It means there's no single best structure; the right one depends on factors like the project's technology, strategy, culture, and how well contributors work together. This suggests that the success of OSS projects is influenced by factors such as how contributors communicate, their experience levels, and the programming languages they use.

1.2 Goal

Our research focuses on pinpointing indicators that forecast the success of Open Source Software (OSS) projects during their crucial incubation period. We aim to explore initial activity patterns, the impact of contribution diversity, and the significance of communication within OSS communities. These insights are intended to guide OSS project strategists in enhancing project success and sustainability.

1.3 Research Question

RQ 1: Can the initial activity patterns in projects predict their long-term success?
RQ 2: Does the diversity of contributions correlate with project sustainability?
RQ 3: How do communication patterns affect project outcomes?

2 Background

In our exploration of Open Source Software (OSS) project sustainability, particularly within the Apache Software Foundation Incubator (ASFI), several key studies provide valuable insights. Discussing Code Quality and Process Metrics [1] highlights the importance of contributions from both minor and major contributors in OSS projects for sustainability, underscoring the need for a diverse range of contributions. Similarly, emphasizing the impact of early participation factors on OSS project sustainability [2], stresses the significance of fostering a vibrant community in the initial stages of development. Examining the inefficiencies associated with forking-based development in OSS [3],identifies modularity, centralized management, and handling hard forks as crucial factors for project success. Showcasing ASF's mature governance and support in propelling projects towards sustainability [4], highlights its role in nurturing early-stage projects. Lastly, investigating forking and pull request practices within OSS communities [5], offers best practices for maintainers to enhance contribution efficiency and foster community cohesion. Collectively, these studies offer a comprehensive understanding of the dynamics in OSS project development, emphasizing early community engagement, efficient development practices, and the role of incubators like ASF in fostering project success.

3 Approach

3.1 Data

The Apache Software Foundation Incubator (ASFI) dataset provides an in-depth look at open-source software development, featuring data on contributors, project statuses, code changes, and communications [6]. It includes:

People Dataset: Details on project contributors.

Lists Dataset: Overview of project statuses and timelines.

Aliases Dataset: Resolves contributor identities.

Filelist Dataset: Tracks changes to the project's files.

Commits Dataset: Records details of code commits.

Messages Dataset: Captures email communications between contributors.

This comprehensive dataset aids in understanding the dynamics of OSS project development and success factors

3.1.1 RQ1: Data for initial activity patterns

The comprehensive dataset from the Apache Software Foundation Incubator (ASFI) is a robust source for understanding the dynamics behind project sustainability. The data encompasses records from 269 projects that have traversed the incubation path, marked by two distinct outcomes: "graduated" indicating a successful transition towards sustainability and "retired" denoting unsuccessful attempts.

Key metrics gleaned from the dataset include:

Incubation time: Duration in incubation before graduation or retirement.

Number of commits: Total commit count, indicating development activity.

Number of committers: Unique contributors committing to the project.

Number of emails: Indicator of communication intensity among developers.

Number of email senders: Distinct individuals initiating email communications, indicating engagement.

Number of files: Count of files in the project, suggesting codebase scale.

3.1.2 RQ2: Data for Diversity of Contributions

The commits table provides data on file operations (additions, modifications, deletions) and the diversity of the affected files. The temporal element of the columns in the Commits table shows the frequency of these operations over time. Data features from the commits table including file_operation, file_name, addlines, and dellines will be instrumental in assessing the diversity of contributions. These features will enable us to analyze the types and extent of changes made to project files, offering insights into the variety of contributions. From the list table, listid, pj_alias, and status will help correlate these contributions with project outcomes (graduated, retired), providing a comprehensive view of how diverse contributions influence project sustainability. Further, the lifespan date range, derived from the start_date and end_date in the list table, will be utilized for regularizing commit contributions.

3.1.3 RQ3: Data for communication patterns

For this research question, the analysis will use two tables:

Messages: communication patterns, commit messages and general discussions.

Lists: information on project statuses (graduated, retired, or in incubation).

The combined dataset resulting from the aggregation and merging process includes the following key columns for each project:

Project ID and Name: Identifiers and names for tracking each project. *Status:* The current status of the project (Graduated, Retired, or In Incubation).

General Discussions: The number of messages not related to code commits, representing unstructured communication.

Commit-Related Messages: The number of messages related to code commits, representing structured communication.

This combined dataset allows for the analysis of how communication patterns, in terms of volume and nature of messages, relate to project outcomes as indicated by their statuses.

3.2 Methodology

3.2.1 RQ1: Methodology

Data Preprocessing

Load commits, filelist, and lists datasets into Pandas DataFrames. Address inconsistencies, duplicates, and missing data. Ensure uniformity in email addresses and developer identifiers. Convert 'commit-datetime', 'start-date', and 'end-date' fields to datetime objects for temporal analysis. Manage missing values through imputation, exclusion, or alternative methods.

Initial Activity Analysis

Identify the first three months post-'start-date' as the crucial window for initial project activity, capturing the essence of early contributions and dynamics. Calculate key metrics within this window—total commits, unique contributors, files added, and files modified—utilizing the 'commits' and 'filelist' datasets. Provide descriptive statistics for these metrics to assess the variability and distribution of initial project activities.

Correlation with Project Outcomes

Data Integration: Merge initial activity data with final project outcomes using 'listid' to classify projects as 'graduated', 'retired', or 'in incubation'.

Statistical Analysis and Modeling

Regression & Evaluation: Use logistic regression to analyze the link between initial activities and project outcomes. Incorporate interaction terms based on exploratory findings and analyze coefficients as odds ratios. Evaluate model performance with accuracy, precision, recall, and F1-score from a classification report and confusion matrix.

Addressing Multicollinearity: Check for multicollinearity with the Variance Inflation Factor (VIF), adjusting the model by removing or merging correlated predictors. Verify the model's fit with the Hosmer-Lemeshow test, ensuring the model adequately represents observed data.

3.2.2 RQ 2

Data Extraction and Preprocessing FExtract commit operations (additions, modifications, deletions), their frequencies, and file name diversity to quantify contribution diversity. Use the file lists table to contextualize file changes and identify changes over the project lifecycle. The lists table also provides project status information (graduated vs. retired) for categorization based on sustainability.

Exploratory Data Analysis Remove outliers based on project start and end times for accurate commit analysis. Analyze commit activity patterns over time using bimonthly bins for ongoing, graduated, and retired projects. Correct for over-counting in multi-file commits to understand unique development efforts. Classify projects by status and visualize their active durations. Create box

plots for commit volumes and operation-specific line change counts, contrasting graduated and retired projects.

Predictive Modeling Use commit and file operation features, along with project outcome, to train machine learning models for predicting project sustainability. Models like logistic regression, random forests, or XGBoost can be employed. Conduct feature importance analysis to identify the most predictive aspects of contribution diversity for project sustainability.

3.2.3 RQ 3

Data Processing Organized, cleaned, and normalized communication data from ASF Incubator projects, categorizing it by project status to enable thorough analysis of communication patterns across different project phases.

Descriptive Analysis Analyzed the average communication messages in different project phases—graduated, incubated, and decommissioned. By calculating these averages, we identified communication trends to set the stage for deeper statistical analysis, aiming to understand how these trends affect project outcomes.

Statistical Testing (ANOVA) ANOVA was applied to assess if communication differences across project statuses were significant. This involved setting up the model, checking assumptions, conducting the test, and interpreting p-values, to ascertain if communication variations are meaningful or random.

Correlation Analysis: We explored the connection between communication and project outcomes. Using the Modularity Index and a correlation matrix, we measured the relationship's strength and direction, aiming to uncover how communication changes influence project success.

Regression Analysis In the regression analysis, we aimed to understand the predictive relationship between coordination message counts and project outcomes. The purpose was to determine the extent to which coordination messages serve as a predictor of project success to gain a deeper understanding of the impact of specific communication types on projects within the ASF incubator.

Variance Inflation Factor (VIF) Analysis Performed VIF analysis to verify our regression model's reliability, checking for multicollinearity among predictors. This ensured our model's robustness, confirming that our findings on communication's impact on project success are valid and reliable.

4 Progress and Accomplishements

4.1 Milestone

Data Acquisition and Cleanup Secured and pre-processed essential ASF Incubator datasets, ensuring data integrity for analysis.

Initial Activity Window Analysis Identified and analyzed the first three months of project activities, calculating critical metrics like total commits and unique contributors.

Outcome Correlation Study Merged activity metrics with project outcomes, utilizing visual explorations to highlight correlations between early activities and project success.

Statistical Modeling Developed a logistic regression model to probe the relationship between initial project activities and outcomes, incorporating checks for multicollinearity and model fit.

Contribution Diversity Analysis Analyzed file operations data to assess the impact of contribution diversity on project sustainability, employing predictive modeling to forecast outcomes.

Communication Pattern Exploration Investigated the role of communication patterns in project success, conducting descriptive, ANOVA, and regression analyses to elucidate the impact of coordination messages.

These milestones mark substantial progress in our understanding of the determinants of project success within the ASF Incubator, setting the stage for deeper insights and strategies to foster open-source project sustainability.

4.2 Result

4.2.1 RQ 1

The analysis indicates that successful projects tend to have higher medians of total commits and unique contributors, with more outliers suggesting active early development and diverse involvement. Conversely, retired projects exhibit lower medians in these metrics, implying less initial activity. For files added and modified, graduated projects also show higher medians, indicating a more dynamic and evolving initial codebase compared to retired projects. Overall, robust initial activity correlates with project graduation, although outliers suggest this is not a strict rule as shown in Figure 1.

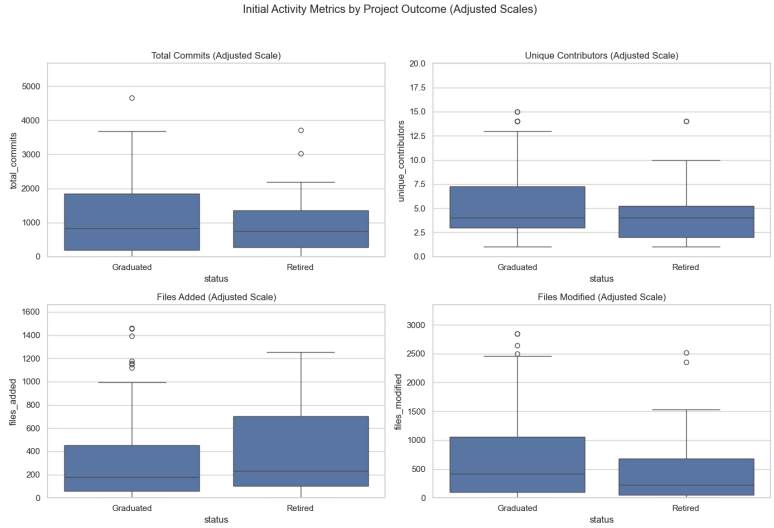

Figure 1: Analysis of Initial Activity Metrics by Project Outcome

The statistical analysis revealed weak correlations between project activities and graduation or retirement outcomes. Total commits and unique contributors showed weak positive correlations with graduation outcomes, but these were not statistically significant (correlation coefficient = 0.0698 and 0.1071, p-values = 0.3807 and 0.1777, respectively). Files added displayed a weak negative correlation with graduation outcomes (correlation coefficient = -0.1025, p-value = 0.1971), while files modified showed a weak positive correlation (correlation coefficient = 0.1146, p-value = 0.1491). These findings suggest that while there may be slight tendencies, these factors alone are not sufficient predictors of project graduation or retirement. The logistic regression model achieved an accuracy of 81%, indicating moderate predictive performance.

Projects with more commits, unique contributors, and a higher number of files added and modified in the initial phase tend to be associated with successful graduation from the ASFI. These findings can guide future project strategies to focus on robust initial development and engagement for long-term success.

4.2.2 RQ 2

These box plots in Figure 2 and Figure 3, respectively, show bimonthly commits of graduated and retired projects across the study interval. It is visually apparent that graduated projects commit more than retired projects across the entire interval. For example, the quartile range of commits around 2007 for retired projects hovered around 10, an order of magnitude below that of graduated projects, which hovered around 100.

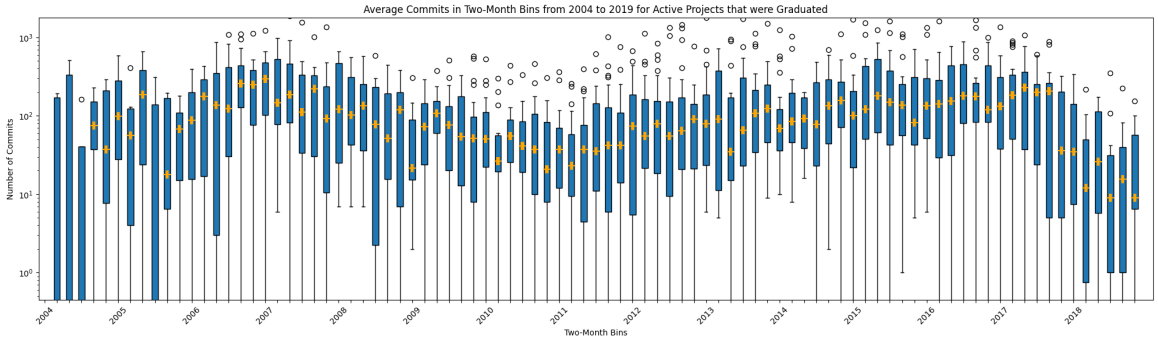

Figure 2: Commit count for graduated projects over time

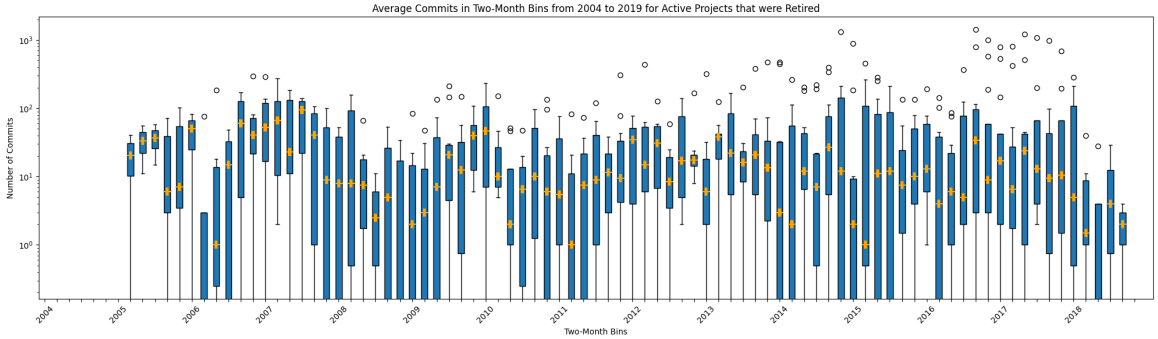

Figure 3: Commit count for graduated projects over time

These box plots in figure 4 show the annual operation-specific line change counts for both graduated and retired plots. We demonstrate that compared to graduated projects, retired projects have an order of magnitude fewer addlines from the add operation. There is also a much higher average and upper interquartile range for two categories, addlines from the copy operation and line deletions from the add operation. There is only a slight overlap around $10\hat{5}$ between the quartile ranges for graduated and retired projects in regard to addlines from the add operation.

4.2.3 RQ 3

At the ASF Incubator, graduated projects had averages of 3997.88 general discussions and 1736.94 commit messages, incubated projects had 4581.46 and 951.89 respectively, and retired projects had 3749.70 and 1166.28 as shown in Figure 5 and 6. Despite these differences, ANOVA results indicate no significant variation in communication volumes across project stages, suggesting that the amount of communication alone doesn't predict project success.

There's a slight negative correlation between coordination messages and project success, with incubation projects showing the highest coordination message counts. This suggests more coordination

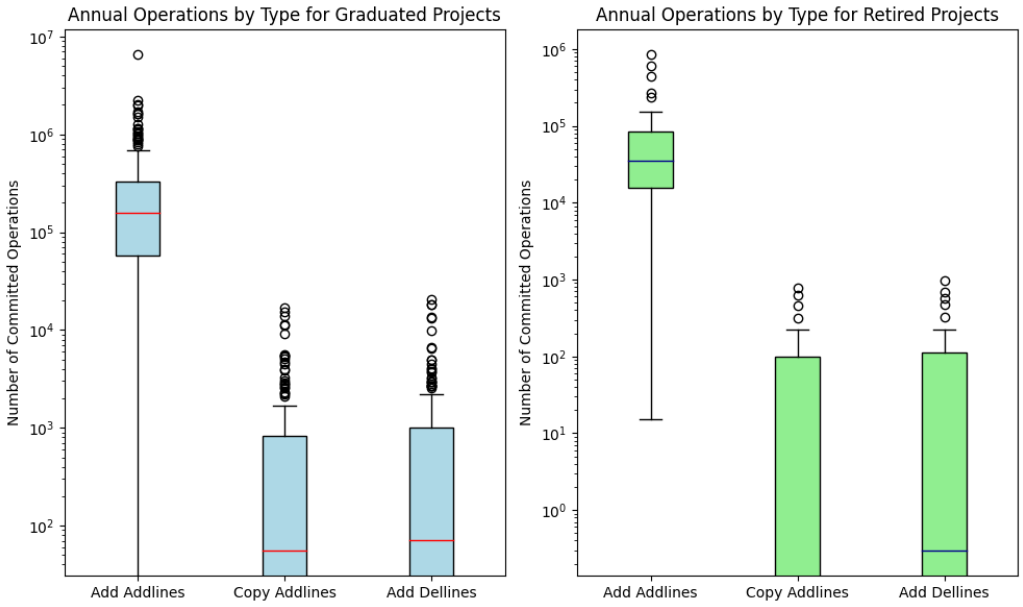

Figure 4: Annual addlines from add, addlines from copy, and dellines from add operations.

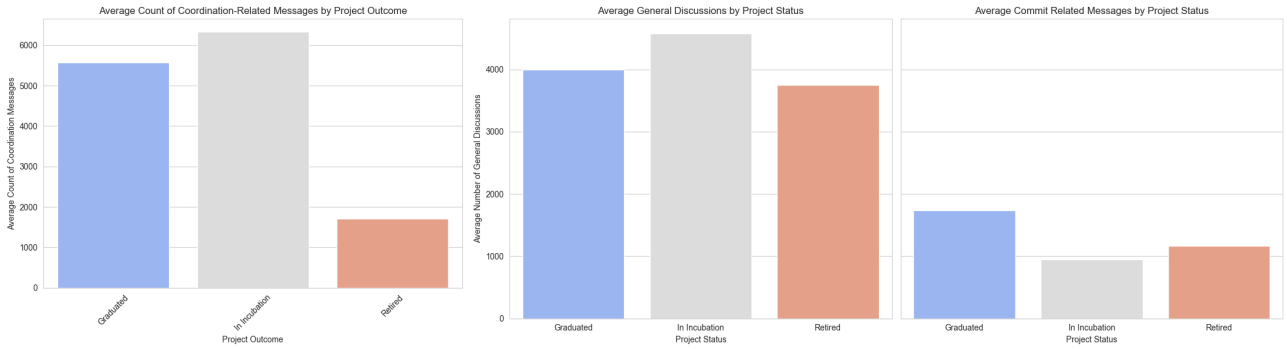

Figure 5: Average Number of General Discussions

isn't necessarily beneficial for project success. The analysis also probes if projects with more commit messages versus general discussions have different outcomes, exploring the effect of communication type on success.

The regression analysis identifies a significant negative relationship between coordination messages and project outcomes, indicating that increased coordination doesn't assure success. Although not linked directly to forking practices, this insight into communication patterns could shed light on how forking impacts project dynamics. The VIF analysis confirms the predictors' independence, validating the regression model's reliability and the robustness of the findings.

4.3 Future work and Timeline

4.3.1 RQ1

Residual Analysis Check for patterns indicating assumption breaches, like non-constant variance or non-normality.

Expansion of Predictor Variables Include additional metrics and contextual factors (e.g.,

nature of contributions, communication patterns, external dependencies) for better understanding and prediction.

Longitudinal Data Analysis Use longitudinal data to track project evolution over time, identifying critical phases and early indicators of success or failure.

Advanced Statistical Modeling Explore sophisticated models (e.g., random forests, support vector machines, neural networks) to uncover complex relationships not captured by logistic regression.

Model Refinement Fine-tune the logistic regression model based on performance and residual insights, and consider alternative predictive approaches for better accuracy and interpretability.

4.3.2 RQ 2

To explore further into the effect of commit diversity on sustainability, we need to perform two more steps.

Predictive Modeling: We need to train machine learning models using the features derived from commits and file operations, coupled with the projects' sustainability status. We will test models such as logistic regression, random forests, or XGBoost, and use the best as our final result.

Feature Importance Analysis: For the feature importance analysis in predictive modeling for RQ 2, we plan to test techniques such as permutation feature importance, which is model agnostic, or else model-specific methods like Gini importance for random forests or SHAP (SHapley Additive explanations) values for models like XGBoost. These techniques will help identify and quantify the contribution of each feature towards predicting project sustainability, offering insights into which aspects of contribution diversity significantly impact the likelihood of a project's success within the ASF Incubator.

4.3.3 RQ 3

Detailed Impact of Forking on Communication Dynamics Utilize data mining to extract communication data pre- and post-forking from project repositories and platforms. Use statistical methods to analyze changes in communication volume and type. Apply network analysis to assess structural changes in communication networks after a fork.

Correlation Between Specific Communication Types and Project Phases Employ NLP to classify communication into categories like problem-solving or decision-making. Use time-series analysis to link communication types with project phases. Integrate machine learning to enhance the accuracy of categorization and correlation.

In-Depth Analysis of Negative Correlation Between Coordination Messages and Project Outcomes Use NLP for content analysis to understand coordination messages. Apply correlation and regression to study the relationship between coordination message characteristics and project outcomes. Use case studies for detailed examination of how coordination affects project success.

5 Progress and Accomplishements

5.1 Critical Immediate Goals in 1 Week

In the next week we will focus on Longitudinal Data Analysis in RQ 1 to identify early indicators of success and failure. We will also explore Predictive Modeling for RQ 2, as this is necessary for properly analyzing feature importance. For RQ 3, we will use Spearman's Rank to quantify the correlation between communication types, such as bug reports, and project outcomes as shown in Table 1.

Research Question	Activity	Start Week	End Week
RQ 1	Residual Analysis & Expansion	8	9
	Longitudinal Analysis(Immediate)	8	9
	Model Refinement	9	10
RQ 2	Predictive Modeling (Immediate)	8	9
	Feature Importance Analysis	9	10
RQ 3	Detailed Impact Data Mining	8	9
	Communication Correlation (Immediate)	9	10
	Proofreading & Final Submission	10	10

Table 1: Remaining Schedule

6 Team Membership and Attestation

Team members Karamjeet Singh Gulati, Bobby Missirian, Pu Sun participated sufficiently.

References

[1] undefinedtefan Stănciulescu, Likang Yin, and Vladimir Filkov. Code, quality, and process metrics in graduated and retired asfi projects. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2022, page 495–506, New York, NY, USA, 2022. Association for Computing Machinery.

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

[3] Anurag Dhasmana, Arindaam Roy, Divjeet Singh Jas, Kiranpreet Kaur, and Pinn Prugsanapan. Forking around: Correlation of forking practices with the success of a project, 2021.

[4] 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.

[5] 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.

[6] 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.