The impact of public investment on open-source digital infrastructure ecosystems hosted on GitHub

Pre-Analysis Plan & Data Report

Paul Sharratt

January 31, 2024

Summary

This research project aims to explore the impact of public investment, specifically from the Sovereign Tech Fund, on open-source digital infrastructure ecosystems hosted on GitHub. The focus lies in understanding how contracted software development payments affect GitHub activity and community health metrics within these ecosystems. The central research question is: "How does public investment impact development activity and community health of open-source digital infrastructure ecosystems on GitHub?"

To conduct this analysis, I will use data from three sources: investment data from the Sovereign Tech Fund, GitHub activity data obtained through GitHub's REST API or the CHAOSS (Community Health Analytics in Open Source Software) project's data collection and visualistion tool Augur, and community health metrics developed by CHAOSS. Additionally, open-source security measurements from the Open Source Security Foundation may be included. Currently, the proposed research design consists of (naïve) longitudinal analyses and a Difference-in-Differences (DiD) methodology.

The project is conducted in partnership with the Sovereign Tech Fund, supported by SPRIND and the Federal Ministry for Economic Affairs and Climate Action of Germany, who have allowed me to use their investment data, and with additional support from researchers at the CHAOSS project of the Linux Foundation. I hope that the findings provide useful insights for policymakers, maintainers, and funders working in open source.

Motivation & Background

This project interests me because it involves an overlap of quantitative data analysis, digital policy, digital public infrastructure, and open-source software development. More broadly, I am especially interested in the role that open source plays and will play in European digital sovereignty initiatives. As open-source software becomes increasingly integral to daily life, government administration, and essential public services, understanding its funding is crucial. My study aims to explore how public investment ultimately affects software health and resilience, contributing to effective and sustainable state funding of open-source development.

My role at the Sovereign Tech Fund/SPRIND GmbH, which invests in open-source projects in the public interest, drives my research. After graduating from the Hertie, I intend to continue working at the Sovereign Tech Fund. Finally, this project is a chance to improve my data science skills, especially in advanced statistical methods (which are lacking because I didn't take Statistics II), and in languages like Python, R, and SQL. Ideally, I would like to collect my data using Python and SQL and conduct my statistical analyses in R.

Introduction

Context: This project is rooted in the evolving landscape of digital infrastructure and policy, particularly focusing on the burgeoning role of open-source software. Digital infrastructure is increasingly dependent on the development of open-source software, a trend underscored by Nadia Eghbal's report for the Ford Foundation, "Roads and Bridges," which emphasizes the growing importance and the simultaneous lack of institutional support for public digital infrastructure and is generally a key text for this policy field.

Key Papers: There are two key papers that are particularly useful for this project. The first, Open Source Community Health: Analytical Metrics and Their Corresponding Narratives by Goggins et al., investigates open source project health, integrating field research and metric development. This study, using CHAOSS project data, argues that the quantification of software developer behaviour and contributions gives rise to particular narratives or ways of interpreting the health of open-source ecosystems. The paper proposes a mixed methods approach for the assessment of project health. However, it does also propose and articulate quantified metrics that may be useful for the assessment of ecosystem health. Within the scope of this thesis, these metrics (such as responsiveness) provide useful outcome variables, assuming they can be operationalized.

The second paper, "GitHub Sponsors: Exploring a New Way to Contribute to Open Source," by by N Shimada et al. examines the GitHub Sponsors program, focusing on its impact on developer activity and community dynamics. It employs a mix of data analysis and surveys, providing empirical evidence on the influence of financial investments in open-source communities, relevant to my research on public investment in these ecosystems. The descriptive approaches in this paper are particularly useful for mapping out the contributions made by sponsored and non-sponsored developers to given organisations and repositories. Additionally, the analysis of the difference in behaviour between sponsored and non-sponsored developers gave me the idea of using a Difference-in-Differences (DiD) methodology for the overall project. Together, these papers guide my approach to understanding the dynamics of open source ecosystems and the role of financial support.

Ethical Considerations: Finally, given that this project concerns open-source software, it is important to highlight the particular ethical components of research in this field. In this regard, Amanda Casari, Julia Ferraioli, and Juniper Lovato's paper Beyond the Repository will be a key reference point for this research project.

Research Question

The central research question for this project is: How does public investment impact development activity and community health of open-source digital infrastructure ecosystems on GitHub?

Hypotheses: In order to operationalize this research question, there are a number of non-directional hypotheses that I wish to test:

1. that funding from the Sovereign Tech Fund is associated with changes in activity on GitHub (such as commits, pushes, code additions & deletions) in specific repositories of software ecosystems
2. that funding from the Sovereign Tech Fund is associated with changes in measures of community health (such as the number and make-up of contributors on GitHub) of a given open-source software ecosystem
3. that funding from the Sovereign Tech Fund is associated with changes in measures of software health (such as responsiveness to issues on GitHub) of a given open-source software ecosystem
4. that funding from the Sovereign Tech Fund is associated with changes in measures of the security (such as the OSSF Security Score) of a given open-source software ecosystem hosted on GitHub

In the course of the thesis, I hope also to make inferences about the kind of metrics that would be useful for analyzing the specific subset of open-source projects that are funded and supported by the Sovereign Tech Fund.

Data & Methods

Data

1. **Investment Data (Dataset 1)**: This dataset comprises financial contributions from an open-source funding organization to GitHub-hosted open-source software projects. It includes the names of recipient organizations, contract duration (start and end dates), contract amounts, specific repositories involved, and contract types.
2. **GitHub Activity Data (Datasets 2 & 3)**: These datasets contain information about the GitHub activities of the funded organizations and data about the specific repositories that constitute an organisation or project on GitHub. They contain repository names, monthly counts of commits and pushes, the number of contributors, code additions and deletions per month, and other GitHub activity metrics.
3. **Community Health Metrics (Dataset 4)**: This dataset measures the community health of open-source software projects. It contains names of the projects and metrics of responsiveness (time taken to respond to open issues) per month.

The analysis will treat Dataset 1 as the independent variable to explore its influence on the variables in Datasets 2, 3, and, if possible, 4. Please note that as of submission, I am still collecting and wrangling datasets 2, 3, and 4.

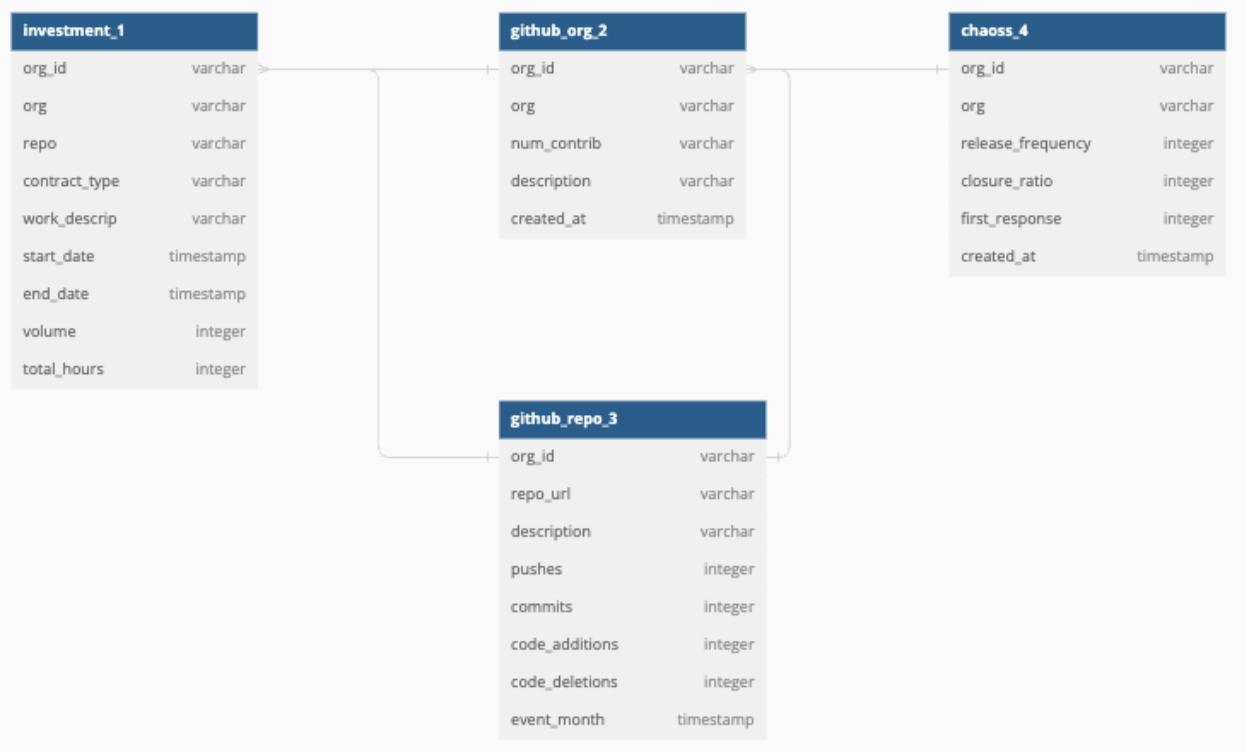


Figure 1: Target Data Schema

Methods

For my initial analysis, I will utilize a longitudinal analysis approach with the datasets. This method will allow me to observe how the investments detailed in Dataset 1 affect the GitHub activities on the organisation and repository level (Datasets 2 and 3) and community health metrics (Dataset 4) over time. By analyzing changes within the same projects before and after receiving funding, and tracking these changes monthly, I can identify trends and potential causal links between funding and project development and health. However, there are potential challenges with the longitudinal analysis approach. These include time-related confounding, as changes observed in the OSS projects might be influenced by external factors beyond funding. Additionally, the evolution of projects due to factors like changes in governance or changes in maintainers and contributors might also impact the results, posing a challenge to isolating the effects of funding. Furthermore, findings from this analysis may have limited generalizability to other OSS projects.

I initially considered a Regression Discontinuity Design, but have decided against it. This is primarily because the funding allocation by the Sovereign Tech Fund doesn't seem to follow a clear, quantifiable threshold, which is key for RDD. The small number of projects, and more generally small sample, around any potential threshold could also weaken the statistical power of the analysis. Moreover, the findings from RDD might lack generalizability beyond the immediate vicinity of the cut-off. Lastly, correctly implementing RDD in this context could be very complex.

An alternative approach could involve implementing a Difference-in-Differences (DiD) methodology. This would entail comparing open-source projects that applied for but did not receive funding from the Sovereign Tech Fund (acting as a control group) with those that did receive funding. For this analysis, however, I would have to identify projects that applied but did not receive funding from the Sovereign Tech Fund and then further collect GitHub activity data using either the CHAOSS Augur tool or the GitHub REST API. By analyzing pre- and post-funding data from both groups, I could isolate and measure the specific impact of funding on variables like GitHub activity. This comparative approach could provide more robust insights into the assumed causal effects of financial support.

Data Report

Summary

During the data collection process so far, I've encountered several challenges that affected the progress of creating, collecting, and wrangling the required datasets. Notably, datasets 2, 3, and 4 pose considerable difficulties. While I've made some progress with datasets 2 and 3, the overall data collection process proved slow. This was largely due to the need to pull data from the GitHub REST API using the Postman API manager, which, unfortunately, I'm not very confident with yet. Additionally, there were issues of data completeness with Google's BigQuery GitHub Activity dataset, which meant I couldn't use it as an alternative. Similarly, OSS Insight, which provides useful visualizations and example API calls, doesn't allow users to export data. On a more positive note, Dawn Foster and Sean Goggins at CHAOSS granted me access to the Augur tool, with which I can create a database of GitHub trace data, covering all required GitHub activity data, on both the organisation and repository levels. However, this approach requires a PostgreSQL instance to host and manage the data constituting datasets 2 and 3 and I am currently working setting that up and integrating it with the Augur tool. Furthermore, the completion of dataset 4 is contingent upon having complete data for dataset 2, and I'm uncertain about the accessibility of historical data and the potential need to rewrite functions created by the CHAOSS team, which could be a highly time-intensive process.

Data collection procedures

1. Collection Procedure for Investment Data (Dataset 1)

I compiled this dataset by reviewing contracts between the Sovereign Tech Fund and funded projects. Out of 33 projects funded by the STF, 27 are hosted on GitHub. I extracted the following information and added it to a Excel sheet:

Table 1: Head of Investment Data (Dataset 1)

org	repo	name	volume	start_date	end_date
https://github.com/GNOME/	https://github.com/GNOME/libxml2/	aevum GmbH (libxml2)	136000	August 3, 2023	December 31, 2024
https://github.com/GStreamer/	https://github.com/GStreamer/qt-gstreamer	Centricular Ltd (Gstreamer RTP)	203000	October 2, 2023	December 31, 2023
https://github.com/systemd/	https://github.com/systemd/systemd/	Codethink (systemd)	455000	September 14, 2023	June 30, 2024
https://github.com/Lullabot/	https://github.com/Lullabot/drupal9ci ;	Drupal Association (Drupal (2023))	278700	October 2, 2023	December 31, 2023
https://github.com/fortran-lang/	https://github.com/fortran-lang/fpm	NumFocus (Fortran)	200000	November 10, 2022	June 10, 2023
https://github.com/apache/	https://github.com/apache/logging-log4j2/	Grobmeier Solutions (Log4j)	596160	August 11, 2023	December 31, 2024

2. Collection Procedures for GitHub Activity Data (Datasets 2 & 3)

The 27 organisations or projects that are on GitHub account between for 473 repositories. In order to collect activity trace data, I have tried two approaches.

Approach 1 - Brute Force: In this approach, I've used the Postman API manager to make individual queries to the GitHub API for retrieving data on repositories associated with the specific projects that STF funds. The process involves several steps:

1. **API Queries**: I construct queries using Postman to access GitHub API endpoints that provide the desired data. These queries include parameters such as organization names, repository names, date ranges, and specific metrics to collect.
2. **Iterative Querying**: I used the API manager to iteratively execute these queries for each open-source project of interest. This involves specifying the project or organization to focus on and retrieving data for each one individually. For org-level data, queries are aimed at collecting aggregate statistics/total data for the entire organization, while for repo-level data, queries are tailored to specific development activities on the repositories.
3. **Data Extraction**: I collected the API responses and extracted relevant data fields. This may involve parsing JSON responses to extract information such as commit counts, code additions and deletions, contributor details, and other GitHub activity metrics.
4. **Data Compilation in RStudio**: Once I have collected and organized the org-level and repo-level datasets, I imported them into RStudio. In RStudio, these datasets can be combined and merged as needed to create a comprehensive dataset that includes information on both the organizational and repository levels.

Table 2: Example of Organisation Data From GitHub API - Repositories

id	full_name	html_url	description
8332666	fortran-lang/fftpack	https://github.com/fortran-lang/fftpack	Double precision version of fftpack
92602565	fortran-lang/vscode-fortran-support	https://github.com/fortran-lang/vscode-fortran-support	Fortran language support for Visual Studio Code
228099016	fortran-lang/stdlib	https://github.com/fortran-lang/stdlib	Fortran Standard Library
232399114	fortran-lang/stdlib-docs	https://github.com/fortran-lang/stdlib-docs	Documentation for https://github.com/fortran-lang/stdlib
233763778	fortran-lang/fpm	https://github.com/fortran-lang/fpm	Fortran Package Manager (fpm)
254476421	fortran-lang/fortran-lang.org	https://github.com/fortran-lang/fortran-lang.org	(deprecated) Fortran website

Table 3: Example of Fortran FPM Repo Activity Data From GitHub API

event_month	pushes	commits
2020-01-01	17	60
2020-02-01	3	18
2020-03-01	4	22
2020-04-01	9	43
2020-05-01	5	17
2020-06-01	8	40

Approach 2 - Augur (CHAOSS) Tool: In this approach, Augur serves as the primary tool for collecting comprehensive data on the open source projects, focusing on measuring their overall health and sustainability. The data collection process involves several steps:

1. **Clone Augur Repository**: Next, I clone the Augur repository using the command git clone 'https://github.com/chaoss/augur.git'. After that, I create a virtual environment with python3 -m venv path/to/venv and activate it with source path/to/venv/bin/activate.

2. **Set Database Connection**: The next step involves setting up a PostgreSQL instance and creating a database. I ensure to specify the database connection string using the format export AUGUR_DB=postgresql+psycopg2://:@:/.

3. **Install Augur**: Following that, I initiate the installation of Augur by running the installation script with either make install or make install-dev if I intend to develop with Augur. This script handles the installation of Augur's Python library, establishes its schema in the database, and prompts me for GitHub and GitLab keys.

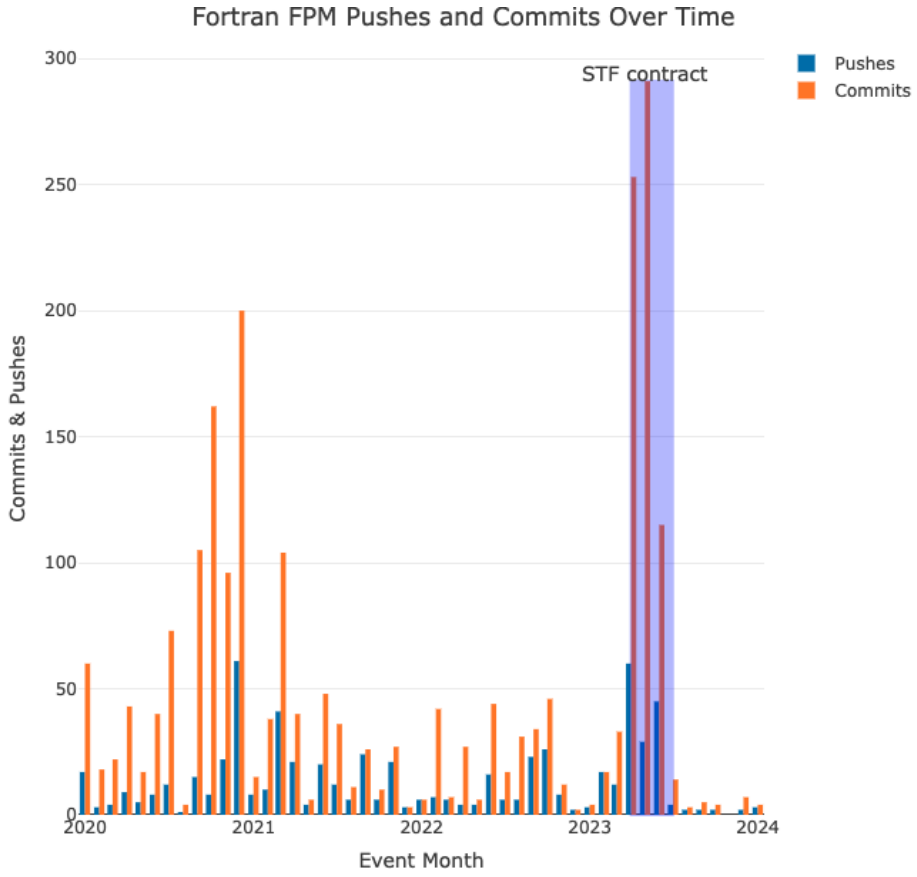
4. **Start Data Collection**: Finally, I'm prepared to commence data collection by configuring Augur's data collection workers tailored to specific data sources. I can refer to the provided documentation to set up these workers effectively, enabling me to gather comprehensive insights into the health and sustainability of open-source projects.

3. Community Health Metrics (Dataset 4)

Dataset 4, the Community Health Metrics, relies on CHAOSS' Starter Project Health Metrics Model. This model draws data from the same Augur database, comprising 473 repositories, to extract the metrics required for the second dataset. The data collection process involves various components, and the model incorporates using the pre-written functions in the to derive the metrics that will in turn constitute the dataset.

Exploratory Data Analysis - Time Series Test

This is an extremely naïve approach, but on the basis of dataset 1 and some initial collection for dataset 3, I can demonstrate a correlation between the Sovereign Tech Fund's funding and activity on a given repository on GitHub. The time series plot below displays the pushes and commits on the Fortran Package Manager repository on GitHub between 2020 and 2024. Overlaid is the period of a contract between the Sovereign Tech Fund and NumFOCUS, which hosts the Fortran community.



As a consequence of the investment contract, NumFOCUS were able to hire three full-time developers to work on the Fortran Package Manager repository to address technical debt and make code improvements. As can be seen in the plot above, during the duration of the investment contract GitHub activity, specifically the number of commits and pushes on the repository, increased considerably. However, whether or not such increases in activity is beneficial for the overall health and sustainability of an open source project is debatable and requires the operationalization of metrics beyond trace data on GitHub. As Goggins et al note, activity often serves as a common indicator for gauging project well-being. However, research that concentrates on activity trace data merely assesses their occurrence without providing robust insights into the "health" of an open-source ecosystem, such as its long-term viability and resilience.

References

1. S. P. Goggins, M. Germonprez and K. Lumbard, "Making Open Source Project Health Transparent," in Computer, vol. 54, no. 8, pp. 104-111, Aug. 2021, https://doi.org/10.1109/MC.2021.3084015.
2. S. Goggins, K. Lumbard and M. Germonprez, "Open Source Community Health: Analytical Metrics and Their Corresponding Narratives," 2021 IEEE/ACM 4th International Workshop on Software Health in Projects, Ecosystems and Communities (SoHeal), Madrid, Spain, 2021, pp. 25-33, https://doi.org/10.1109/SoHeal52568.2021.00010.
3. Casari, A., Ferraioli, J., & Lovato, J. (2023). Beyond the Repository. Communications of the ACM, 66(10), 50–55. https://doi.org/10.1145/3605160.
4. Shimada, N., Xiao, T., Hata, H., Treude, C., & Matsumoto, K. (2022). GitHub Sponsors: Exploring a New Way to Contribute to Open Source. 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE), 1058-1069. https://doi.org/10.48550/arXiv.2202.05751