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

Pre-analysis Plan

Paul Sharratt

31. January 2024

Summary

This project investigates the impact of public investment on open-source software. My objective is to examine the impact of financial support, in the form of contracted software development payments funded by the Sovereign Tech Fund, impacts various key metrics within open-source digital infrastructure ecosystems hosted on GitHub.

Data

In order to answer this question, I am collecting or have collected data from two main sources; contractual data from the Sovereign Tech Fund and GitHub activity using GitHub's REST API or a data collection and visualisation tool called Augur developed by the CHAOSS (Community Health Analytics in Open Source Software). In addition, I intend to utilize selected metrics for assessing opens-source ecosystem health developed by CHAOSS and, if possible, include measurements of open-source security from the Open Source Security Foundation. Currently, I have two methods for collecting data on GitHub organizations and repositories described below. In the course of the next weeks, I expect to settle on one data collection method.

Method

This research employs a quantitative data analysis methodology, relying on publicly available data from GitHub and other relevant sources. The methodology comprises essential stages, including data collection, data cleaning, and preprocessing of GitHub repository data, public investment data, and supplementary repository security data.

To ascertain the relationships between public investment and the open-source health metrics, advanced statistical analyses, I intend to use a regression discontinuity design. This will be applied to investigate the impact of public investment on GitHub activity, community health metrics, and security within select open-source digital infrastructure projects.

Importance

This research holds considerable significance as it advances our understanding of the role of public investment in fostering open-source digital infrastructure projects. It directly aligns with the mission of the Sovereign Tech Fund and benefits from collaboration with researchers from the CHAOSS project at the Linux Foundation. The findings will offer valuable insights for policymakers and organizations like the Sovereign Tech Fund, demonstrating the potential benefits of strategic investments in open digital infrastructure. By addressing this question, I aim to inform evidence-based policy decisions and contribute to the responsible allocation of public resources in the digital age, with the support of key stakeholders in Germany and the open-source community.

Support

The project will be carried out in collaboration with the Sovereign Tech Fund, supported by SPRIND (the Bundesagentur für Sprunginnovationen) and the Federal Ministry for Economic Affairs and Climate Action of Germany, alongside support from researchers from the CHAOSS (Community Health Analytics in Open Source Software) project at the Linux Foundation.

Motivation and Background

The research question of how public investment impacts open-source software ecosystems holds interest for me for several reasons. Firstly, as the digital transition accelerates, more aspects of our daily lives and essential services are becoming reliant on open-source software. Understanding and funding this infrastructure is vital for the sustainable development of society. Increasingly, state actors are engaging in the funding, development, and adoption of open-source software. In the European context, this initiative is taking place under the auspices of the broader digital sovereignty policy paradigm. The possible development of European Public Digital Infrastructure Fund illustrates the increasing attention paid to

Secondly, my interest in this topic is rooted in my current part-time position with the Sovereign Tech Fund. This role has provided me with insights into the intricacies of public investment in technology projects, which is directly relevant to the subject of this research. My prior coursework has given me a solid foundation in statistical analysis, econometrics, and data management, all of which are crucial for conducting empirical research in this domain. After graduating from the Hertie, I hope to work for the Sovereign Tech Fund full-time and continue to work on addressing the research question and other related topics.

Finally, through this research project, I hope to improve my data science skills. Despite never having taken Statistics II, I hope to become proficient in advanced statistical techniques, such as regression discontinuity design and time-series analysis. This research will also deepen my understanding of open-source software development and community dynamics, and improve my Python, R, and SQL skills. Ultimately, this project will equip me with the skills to assess the effectiveness of public investments in technological innovation.

Introduction

The research question is situated within a broader context of digital infrastructure, digital policy, and the growing significance of open source software. Modern digital infrastructure is built with and relies on the development of open source software. For a general overview of the increasing importance of and lack of institutional support for public digital infrastructure, Nadia Eghbal's report for the Ford Foundation, Roads and Bridges, is instructive.

There is a considerable amount of research constructing and presenting indicators of open source project activity, but a lack of consensus about how indicators derived from trace data might be used to represent a coherent view of open source project health and sustainability.

Currently, some 27 organizations that host their code on GitHub have received funding for work from the Sovereign Tech Fund. The duration, volume, and type of work varies from project to project. Between them, these 27 organizations account for 470 repositories on GitHub. These repositories range from simple READ.ME repositories to large and complex repositories with multiple dependencies in a range of programming languages.

The key researcher here is Sean Goggins

Key Papers

Paper 1. - Open Source Community Health: Analytical Metrics and Their Corresponding Narratives

The paper "Open Source Community Health: Analytical Metrics and Their Corresponding Narratives" by Sean Goggins et al. explores the evaluation of open source projects' health and sustainability. It emphasizes the limitations of existing methods that primarily rely on trace data and proposes an approach integrating field research and metric standards development. The paper presents the work of the Linux Foundation's CHAOSS project over four years, highlighting the importance of metrics that incorporate comparison, transparency, trajectory, and visualization. This approach aims to provide a comprehensive understanding of open source software health, connecting trace data with human experience. The relevance to your thesis lies in its focus on assessing open source ecosystem health, an aspect central to understanding the impact of public investment on such projects.

Paper 2. - GitHub Sponsors

The paper "GitHub Sponsors: Exploring a New Way to Contribute to Open Source" investigates the GitHub Sponsors program, examining characteristics of sponsored developers, sponsors, and the effects of sponsorship. It employs mixed methods, including data analysis and surveys, to understand the impact of financial support on open-source developers. The findings suggest that sponsored developers are more active, and the presence of sponsorship influences the community dynamics. This paper's focus on the economic aspects of open-source projects and the impact of financial support aligns closely with your thesis topic, especially in understanding how financial investments can influence open-source software ecosystems.

The quantitative methods used in the "GitHub Sponsors" paper include data analysis of the GitHub Sponsors program. This involves statistical analysis of the characteristics of sponsored developers, sponsors, and their activities on GitHub. The paper utilizes metrics related to the developers' activities and sponsorships, such as the number of sponsors, the frequency of contributions, and the nature of these contributions. This approach helps quantify the impact of financial sponsorship on developer engagement and community dynamics within open source projects.

Similarly, as Goggins et al. note, the demand for coherent, consistent and actionable metrics is growing as a result of state and corporate engagement with and investment in open source. I hope that through this research project, I can get to grips with these debates and, if possible, make a useful contribution.

Ethical Considerations

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:

What is the impact of public investment on open-source digital infrastructure ecosystems hosted on GitHub?

Hypotheses

In order to operationalize this research question, there are a number of 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 specifically the subset of open-source projects that are funded and supported by the Sovereign Tech Fund.

3. Description

Enter a description of your study here.

4. Hypotheses

1. List any *a priori* hypotheses here.

Design Plan

5. Study Type

- Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.
- Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, ñnatural experiments,î and regression discontinuity designs.
- Meta-Analysis - A systematic review of published studies.
- Other (explain your study type)

6. Blinding

- No blinding is involved in this study.
- For studies that involve human subjects, they will not know the treatment group to which they have been assigned.
- Personnel who interact directly with the study subjects (either human or non-human subjects) will not be aware of the assigned treatments. (Commonly known as "double blind").
- Personnel who analyze the data collected from the study are not aware of the treatment applied to any given group.

7. Is there any additional blinding in this study?

Describe any additional blinding procedures here or state not applicable.

8. Study design

Describe your study design here.

9. Randomization

Enter details on randomization here or state not applicable.

Sampling Plan

10. Existing data

- Registration prior to creation of data: As of the date of submission of this research plan for preregistration, the data have not yet been collected, created, or realized.
- Registration prior to any human observation of the data: As of the date of submission, the data exist but have not yet been quantified, constructed, observed, or reported by anyone - including individuals that are not associated with the proposed study. Examples include museum specimens that have not been measured and data that have been collected by non-human collectors and are inaccessible.
- Registration prior to accessing the data: As of the date of submission, the data exist, but have not been accessed by you or your collaborators. Commonly, this includes data that has been collected by another researcher or institution.
- Registration prior to analysis of the data: As of the date of submission, the data exist and you have accessed it, though no analysis has been conducted related to the research plan (including calculation of summary statistics). A common situation for this scenario when a large dataset exists that is used for many different studies over time, or when a data set is randomly split into a sample for exploratory analyses, and the other section of data is reserved for later confirmatory data analysis.
- Registration following analysis of the data: As of the date of submission, you have accessed and analyzed some of the data relevant to the research plan. This includes preliminary analysis of variables, calculation of descriptive statistics, and observation of data distributions. Please see cos.io/prereg for more information.

11. Explanation of existing data

Describe any existing data or state not applicable.

12. Data collection procedures

Enter a description of your data collection procedures here.

13. Sample size

Enter a description of your sample size here.

14. Sample size rationale

Enter your sample size rationale here.

15. Stopping rule

Enter your stopping rule here or state not applicable.

Variables

16. Manipulated variables

Describe manipulated variables here or state not applicable.

17. Measured variables

Describe your measured variables here.

18. Indices

Describe any indices here or state not applicable.

Analysis Plan

19. Statistical models

Describe your planned statistical model(s) here.

20. Transformations

Describe any transformations here or state not applicable.

21. Inference criteria

Describe your inference criteria here or state not applicable.

22. Data exclusion

Describe your data exclusion criteria here or state not applicable.

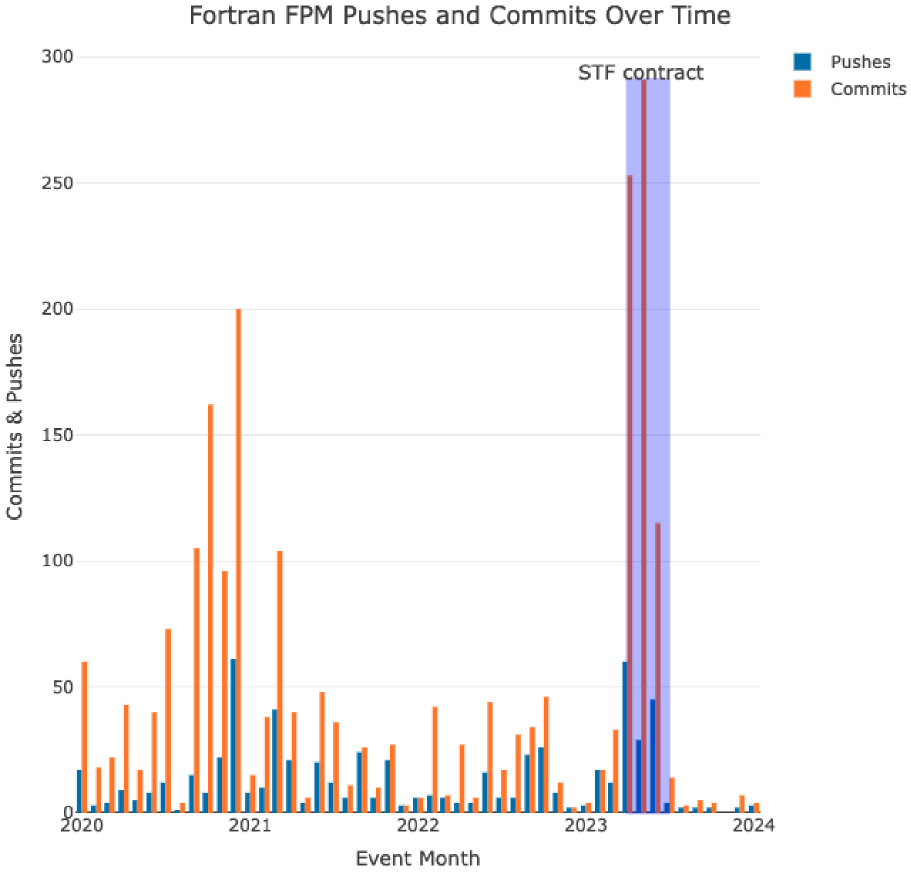
23. Missing data

State your process for dealing with missing data or state not applicable.

Exploratory analysis

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
228099010	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

id	full_name	html_url	description
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



This time series plots 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 substantial insights into the intricate aspects of long-term viability and resilience. “The risk is that numbers without context will aid in the development of stories that are not focused on increasing health and sustainability, but instead focused on telling stories that emphasize activity over health.”

Other

25. Other

Enter any additional information not covered by other sections, or state not applicable.