The Geopolitical Turn in Open Source Software

Jack Lynch, 2025

Abstract

As open source software (OSS) becomes increasingly central to global digital infrastructure, its development and governance are no longer purely technical concerns—they are deeply entangled with geopolitical dynamics. This project explores the evolving geopolitical landscape of OSS by gathering large-scale data on repositories, contributors, and their interconnections across platforms. Using a breadth-first scraping algorithm, we collected structured metadata on projects, user activity, and collaboration networks, with the goal of mapping the broader ecosystem of open source software production.

Introduction

Open source software (OSS) has become a cornerstone of global technological infrastructure, relying on transparent collaboration and free access. Yet, the nature of OSS development may be shifting. In open digital environments, OSS benefits from borderless contribution and visibility. In contrast, countries like China—with highly controlled internet infrastructures—may foster distinct OSS ecosystems. This project aims to establish a base case for the global open source ecosystem, to later be compared to more restricted open source communities, specifically chinese OSS platforms like Gitee, which in fact began to catch up to GitHub in activity starting in 2022.

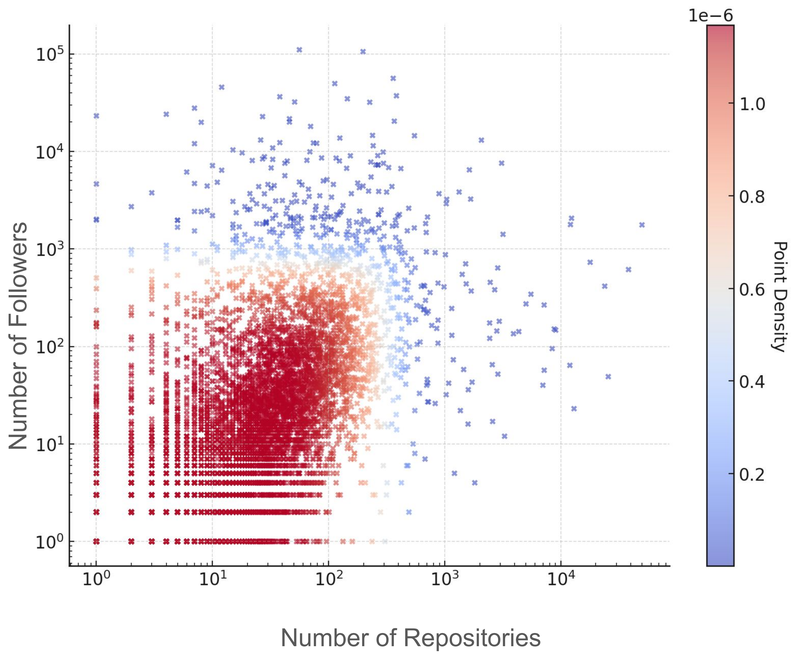
Methods

I implemented a breadth-first scraping algorithm using the GitHub API. Starting from a set of seed repositories, the algorithm traverses repository connections based on forks, stars, and contributor activity to map the network of OSS projects and users.

The data collected for users includes metrics such as starred or followed repositories, followed users, and number of owned repositories. For repositories we tracked similar data for stars and follower counts, but additionally tracked metrics like number of open and closed issues, average issue close time, and whether the repository is a fork (a clone) of another.

While GitHub serves as the primary platform for global OSS, we are also identifying and cataloging Chinese-based alternatives and mirrors to capture region-specific activity not fully visible on global platforms.

Distribution of Users by Repository and Follower Counts



Results

Data collection is currently ongoing, with over 30,000 repositories and contributor profiles successfully scraped from GitHub. This phase of the project focused on building a foundational dataset and conducting preliminary analysis, rather than achieving full coverage of the platform. While most network-related metrics—such as forks, contributors, and repository ownership—were reliably captured, some indicators proved less trustworthy. In particular, star and watcher counts were frequently inflated by bot activity, limiting their analytical value. These early findings underscore the need for careful metric validation and point toward future refinements in data quality and scope.

Conclusions

Once again, this project was not primarily on analysis, but several conclusions can be gained even in this early stage:

- **The majority of users have low activity.** Most users have few repositories and followers overall, with the median repositories per user being around 20, and the median followers 11, and the mode for each is 1 and 0 respectively.
- **A small number of users created the majority of repositories and followers.** The majority of repository content gathered belonged to a small, but prolific group who significantly skew the average (Special congratulations to seanpm2001, who had over 50000 repositories).

While the data gathering is still ongoing on github, which will hopefully reinforce the trends currently seen further, there is still lots of potential for further development. Future work could expanding the sources for the data gathering to other open source hosting platforms such as Azure, Gitlab, and Bitbucket. Another option would be to look more closely at data such as activity over time, which was not gathered in this project.