Estimating Usage Of Open Source Projects

Sophia Vargas
Google
New York, NY, USA
sophiavargas@google.com

Georg J.P. Link
Bitergia
Omaha, NE, USA
georglink@bitergia.com

JaYoung Lee
Google
Mountain View, CA, USA
jayounglee@google.com

ABSTRACT

While open source projects can benefit from usage telemetry to understand their impact, privacy concerns often arise. This paper explores using publicly available metrics as proxies for actual usage data. Using the Flutter project as a case study, we found a strong correlation between these public metrics and Flutter's active user count, demonstrating that such publicly available metrics can offer insights into project usage while respecting user privacy.

KEYWORDS

Open Source Project Usage, Telemetry Data, Project Health Metrics

ACM Reference format:

Sophia Vargas, Georg J.P. Link, and JaYoung Lee. 2024. Estimating Usage Of Open Source Projects. In *21st International Conference on Mining Software Repositories (MSR '24), April 15-16,* Lisbon, Portugal, 2 pages. https://doi.org/10.1145/3643991.3645066

1 Introduction

Usage telemetry can be an effective tool to assess the value and impact of an open source project, and enable time-strapped maintainers to prioritize support for features, versions, and/or components that are actively used. However, many open source users see incorporating usage telemetry into a project as an invasion of privacy, especially for projects that did not launch with embedded telemetry. [1] Even with an adequate value proposition, many users will opt out, or choose not to opt in to share their data with a project, organization, or company. Tools[2] are emerging to address this need for usage telemetry; however, many are still met with skepticism by privacy-minded community members. It is therefore important to have reliable proxy metrics for many projects and companies that need to show the value and impact of open source projects. We are not aware of any research that correlates actual usage telemetry with project metrics, perhaps because very few projects have telemetry data and it is not readily available for researchers.

RQ: What available open source project metrics[3] provide a sufficient proxy for usage telemetry? This proof-of-concept exercise seeks to identify any relationship between usage as measured by embedded telemetry and as indicated by proxy metrics to evaluate whether less-invasive methods are sufficiently reliable for the detection of usage trends in open source projects.

ACM ISBN 979-8-4007-0587-8/24/04.
https://doi.org/10.1145/3643991.3645066

2 Approach

Our study's object is the Flutter[4] open source project, a UI framework released by Google under the BSD-3-Clause license in May 2017. We assembled[5] [6] and evaluated 10 metrics[7] [8] from publicly-available data sources—GitHub[9], StackOverflow, and Slack[10]— to calculate monthly proxy metrics and compare them with Flutter's actual monthly active users (MAU)[11], collected by embedded telemetry during the same time period: Jan 1, 2018 - Feb 28, 2021.

In our test dataset, all of our variables were quantitative with no visible outliers and all proxy metrics had a linear relationship with Flutter MAU demonstrated by r^2=.53 for fork events, and r^2>.71 for all other proxy metrics. As our dataset did not follow a normal distribution—the data is not random and most variables showed variation between mean, mode, and median—we evaluated the relationship between these metrics using both Pearson and Spearman correlations.

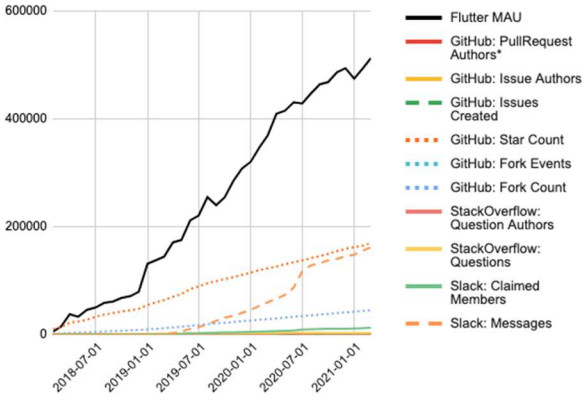

Figure 1: Metrics considered for this analysis[12]

3 Analysis

All proxy metrics considered showed strong positive correlations with MAU, with Pearson results showing r(36)>.73, p<.001 for all proxy metrics, and Spearman results showing r(36)=.66, p<.001

for fork events and r(36)>.81, p<.001 for all other metrics. These results are intuitive, as we expected most of these metrics to increase alongside growth of the project. Comparing month over month (MoM) growth rates between MAU and each proxy metric also produced some positive correlations, with Pearson results showing r(35)>.70, p<.001 for all StackOverflow metrics, cumulative fork count, and GitHub issue authors, and r(35)>.89, p<.001 for star cumulative count and fork events. Spearman results for MoM comparisons yielded r(35)>.61, p<.001 for cumulative fork and star counts. Given the linear nature of MAU (r2=.98), we also attempted to assess our proxy metrics on their ability to forecast actual usage by calculating the slope of (MAU/proxy) and the average across all per month ratios (MAU/proxy), and the relative standard deviation between these values. From this basic approach, the cumulative fork count metric demonstrated the most consistent relationship to actual usage with the relative standard deviation of 4% between the slope and average of these per month ratios.

	Metric	Pearson Results	Spearman Results
Overall	GH Fork count	r(36)=.99, p<.001	r(36)=1.00, p<.001
MoM	GH Fork events	r(35)=.92, p<.001	r(35)=.59, p<.001
Overall	GH Star count	r(36)=.99, p<.001	r(36)=1.00, p<.001
MoM	GH Star count	r(35)=.89, p<.001	r(35)=.61, p<.001

Table 1: Strongest positive correlations

4 Future Work

Our proof of concept analysis found that some open source project health metrics may be suitable proxies for usage metrics. We encourage researchers to continue this exercise with a larger scope across different open source projects and more robust modeling techniques. Our intuition is that open source project health metrics hold up as proxy usage metrics, reducing the need for open source projects to collect usage telemetry.

ACKNOWLEDGMENTS

This work is supported by Google Open Source and Bitergia. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Google Open Source or Bitergia.

ENDNOTES

[1] This blog series discusses Golang's proposal to add telemetry: Russ Cox. 2024. Opting In to Transparent Telemetry. Retrieved from https://research.swtch.com/telemetry-opt-in

[2] Scarf tracks software downloads and usage: Scarf Systems. 2024. Retrieved from https://about.scarf.sh; OpenTelemetry captures and exports metrics, traces, and logs from run-time usage: OpenTelemetry. 2024. Retrieved from https://opentelemetry.io

[3] Georg Link, Sophia Vargas. 2022. 8 ideas for measuring your open source software usage. Retrieved from opensource.com/article/22/12/open-source-usage-metrics

[4] Flutter. see https://Flutter.dev

[5] We used an instance of Bitergia to collect metrics from StackOverflow and GitHub code changes and Issues. Bitergia. 2024. See https://bitergia.com

[6] We referenced the GitHub archive (GHarchive) project via Google BigQuery for historical records of star and fork events. See https://gharchive.org

[7] Proxy metrics evaluated: Slack: # of Claimed Members and cumulative # of Messages posted; GitHub: # of issue authors per month, # of pull request authors* per month, cumulative star count, fork events per month, cumulative fork count; StackOverflow: # of question authors per month, # of questions posted per month. *GitHub pull request authors is the only metric that excludes contributors from Google.

[8] As Google governs changes to this code base, we excluded known Google employees in code related metrics using Bitergia's native functionality that assigns organizations to contributors by email domain to filter out Google contributors. See https://github.com/chaoss/grimoirelab-sortinghat

[9] These metrics included activity in the entire Flutter organization which hosts 32 repositories.

[10] This Slack channel was created in March 2019; for all reported Slack metrics, df=22 for overall and df=21 for MoM

[11] MAU: is calculated as the number of unique users in a 30 day period, reported on the last day of the month. Note that users can opt out at will. It was estimated that less than 2% of new users had opted out for this time period when they were given the choice.

[12] Slack defines Claimed Members is defined as: "The Number of people (not including deactivated members or guests) who have signed in to your Slack workspace or organization at least once" . Workspace administration. 2024. Slack Technologies. Retrieved from https://slack.com/help/articles/360057638533-Understand-the-data-in-your-Slack-analytics-dashboard