Open Source Metrics @Twitter

Developing a better understanding of the health of your community

Remy DeCausemaker @twitter
Casey Haber
Gordon Li
Nyssa Chennubhotla
Miguel Arreguin

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

Overview

The twitter metrics team is focused on the creation of a simple yet powerful tool for assessing the health of open source projects. We are creating a better system for driving community involvement and increase overall code health. Metric dashboards exist for this reason but they are heavy, cluttered and rarely provide the quick facts and data that are needed to drive the push towards a more successful software project. We are building a light-weight and cost-free visualization based report that leverages the github API and can be deployed on github pages as a static javascript bundle. Our report is broken up into a four part narrative: Discovery, Usage, Retention and Activity. Based off of these metric categories and the real time comparison of software projects, our digital narrative will inform and drive a better understanding of your Open Source project.

Project Scope

The purpose of this project is to develop access and understanding of your repository. In order to accomplish this within the duration of a semester, we narrowed the amount of metrics for insight (See Health Categories.) This narrowing of metrics is our solution to the dashboard problem. The scope of this research is to generate the first open source platform for quick and meaningful insights. The tools used and deployment strategies are based off of the open source paradigm in which we want the solution to be simple but easily reprogrammable for future development.

INTEGRATION ANALYSIS CHART

	GRIMOIRE/ CAULDRON.IO	CREGIT	NETFLIX OSS TRACKER	OPENHUB	PROSPECTOR	STACKALYTICS
GIT	✓	✓	✓	✓	✓	✓
GITHUB	✓	✓	✓	✓	✓	✗
PHABRICATOR	✗	✗	✗	✗	✗	✗
JIRA	✓	✗	✗	✗	✓	✗
CONFLUENCE	✓	✗	✗	✗	✗	✗
JENKINS	✓	✗	✗	✗	✗	✗
MAILING LIST	✓	✗	✗	✗	✓	✓

Backend

In order to reflect the nature of open source, the metrics application was built using React.js and webpack to create a static, low-dependency, and zero-cost software project.

Data Collection

We used the GitHub developer APIs for collection of raw repository-level data.This level of granularity was essential for us to drive these simple decisions. The first goal was collection, to write simple reusable scripts that would pull data and do some minimal processing. The next stage was to transfer the functionality to node.js so that it could be easily transferred onto our static application that is deployed using GitHub pages.

Health Categories

1. Retention
2. Maintainer Activity
3. Discovery
4. Usage

Related Work

Open source software development is a heavily invested field with all types of stakeholders. The value of investing in the open source community is continuing to drive huge popularity in tech. The rise of collaboration tools and platforms, such as git, has lead to a boom of “public programming” even causing companies such as Microsoft to contribute to the core of these tools. These stakeholders have caused there to be a gap for the need of collaboration data analysis and understanding. We began this project with a competitive analysis of existing analysis and dashboard platforms (see integration analysis right.) The analysis was correct in confirming that all of the enterprise solutions to open source analysis failed to drive meaningful and digestible insights. While most of them required deployment, storage and more hardware. We decided that in order to push the benefits of open source, we would use only free and open source tools in order to remain transparent and available to people and companies of all sizes.

Platform Development

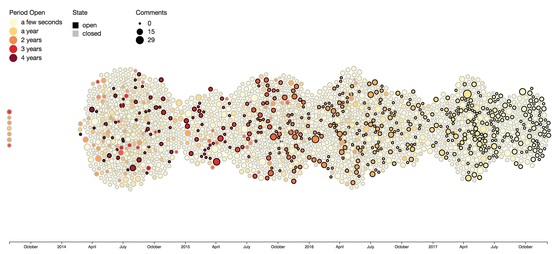
Visual Encoding

Our process for finding the optimal visual encodings began in an exploratory manner. We wanted to maintain granularity as well as pushing the fail or pass portion of the category. After some initial exploration using tableau and semiotic, beeswarm plots best suited our goal of simplistic messaging and high granularity. Similar to frequency plots, they allow for quick insights while maintaining the integrity of the data and is able to maintain data transparency. We split it into four visualizations with a optional fifth heatmap to show what time of day the repo is being worked on (See screenshots of the visuals below.)

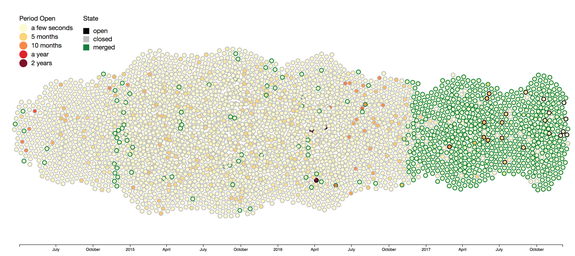
Future Work

This project has immense potential for future collaboration.. Our next goal is to add another iteration of visualizations with annotations and more text. Another aspiration would be to create a similar page but for entire organizations, not just repositories. Beyond that we want to include other open source community tools and activity in our data analysis such as slack, jira, confluence etc,. Special thanks to upstream metrics organizations like CHAOSS and Bitergia for your support and for reviewing our designs and results.

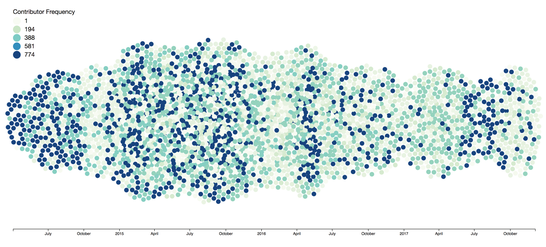
Maintenance: Issues



Maintenance: Pulls



Retention: Pulls



Retention: Commits