MEASURING THE UNIVERSE OF OPEN SOURCE SOFTWARE

Cong Cong, Calvin Isch, Eliza Tobin, Gizem Korkmaz, Bayoan Calderon, Brandon Kramer, and Aaron Schroeder

Social and Decision Analytics Division, Biocomplexity Institute and Initiative, University of Virginia

Project Introduction

- There are no commonly used metrics to measure the amount of open source software (OSS) in existence.
- This project aims to measure how much OSS exists and to better understand the distribution of OSS creation across various sectors to evaluate the economic impact of OSS.
- Traditional measures of innovation looking at the amount of copyright, patents, and trademarks in existence do not accurately capture the universe of OSS innovation. This project defines the OSS universe as all GitHub repositories that have a registered OSI approved license.

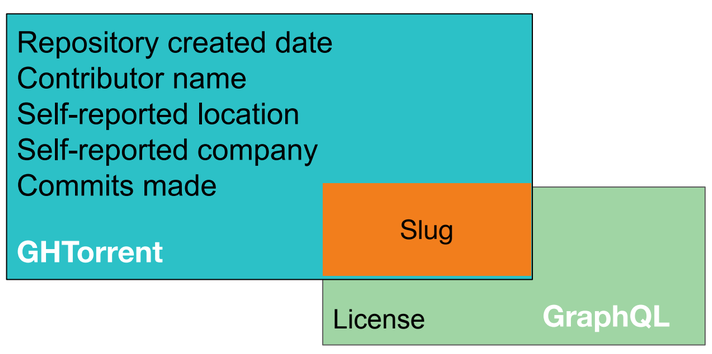
Terminology

Open Source Initiative (OSI) - a worldwide non-profit that spreads knowledge about OSS, promotes its usage, and connect various OSS communities.
Open Source Software (OSS): "a computer software, with its source code made available with a license, in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose" (Source: OSI).
OSS Licenses: Protections defining the limitations of use for code that provide developers with claims when terms are violated. To qualify as an OSS license, a license must allow software to be freely used, modified and shared.

GitHub - world's largest website for developers to build and share software.
Repository - contains all of a project's files
Commit - an individual change to a file for a repository
Contributor - someone who successfully committed to a project (GitHub Glossary)
Slug - URL friendly combination of a username and repo name (*i.e. ulialang/julia)*

Data Collection Methods

- Querying for the OSS licenses, resulted in over 7 million OSI-license holding repositories on GitHub. We used multiple sources to access information on each of these repositories.
- Using **GraphQL,** (GitHubs current API system), we searched for repository names, owners, and license information for each of the OSI approved licenses from 2012-2018.
- These repository names and owners were combined into "slugs" that could be used to identify OSS repositories in an alternative online database, **GHTorrent.**
- Combined we had access to more information on each repository contributors including location, organization, and commit count



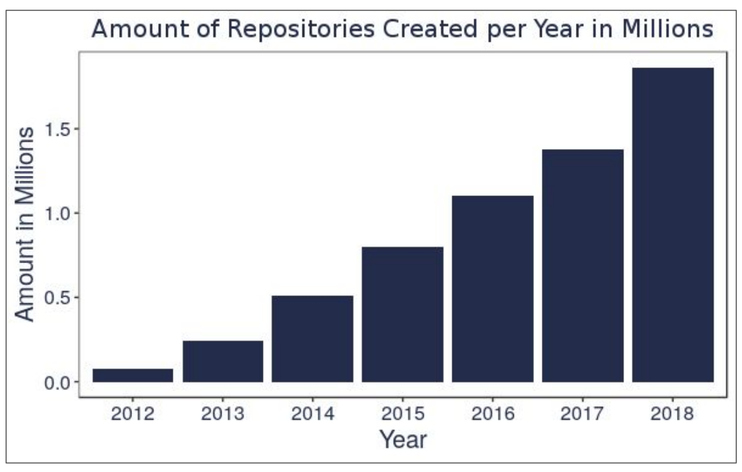
Licenses Expanded

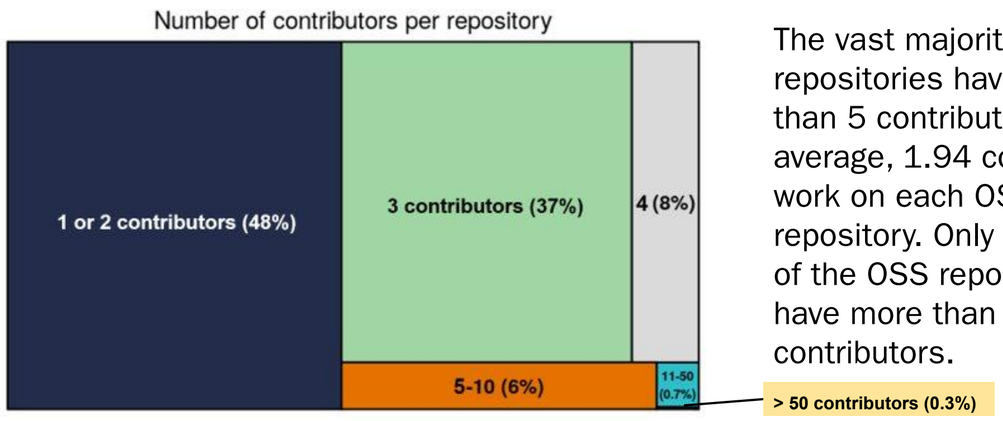
- Licenses provide developers with rights over their own work while promoting the dispersion of free, accessible code.
- License regulations vary. For example, the MIT license allows developers to use the code for any purpose. The GPL license grants individuals with the ability to use the respective code under the stipulation that it remains open source.
- Although 87 OSI-approved licenses exist, the top 13 licenses comprise more than 99% of OSS repos on GitHub.

Data Analysis

From 2012-2018 there were **55,110,925** distinct repositories with commits made on GitHub. Of those, **4,940,788** repositories were Open source–**8.97%** of all projects.

To examine the growth rate of OSS on GitHub, we plotted the number of GitHub Repos with an OSS license each year. The number has been increasing. In 2012 there were 79,400 repos open source licenses. By 2018, there were 1,865,720.

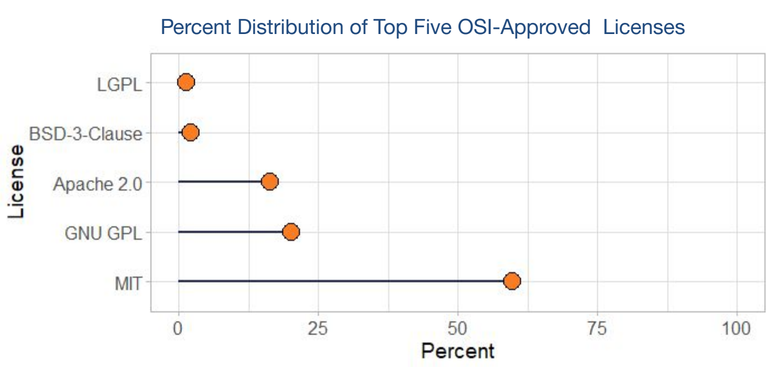




The vast majority of repositories have fewer than 5 contributors. On average, 1.94 contributors work on each OSS repository. Only 0.3% of all of the OSS repositories have more than 50 contributors.

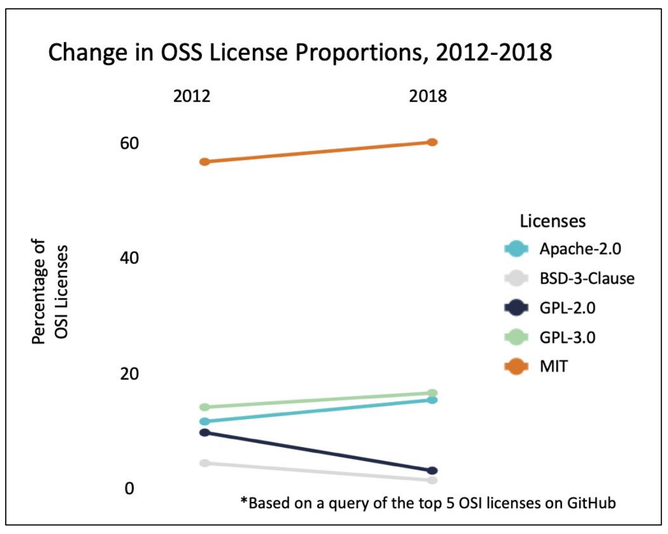
OSS Universe

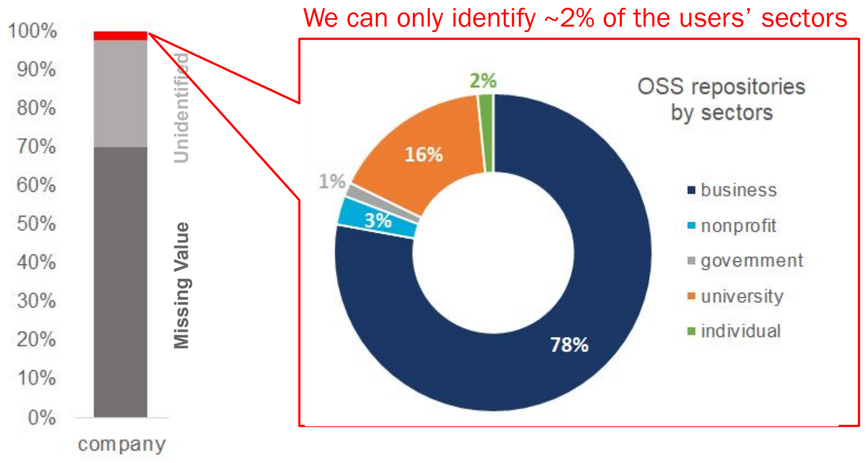
- There are 2,833,618 unique OSS contributors
- OSS grew 2,350% from 2012-18
- On average one repository receives 36.5 commits



(MIT is the most common OSS license with 60%. The next four (GNU GPL, Apache-2.0, BSD-3-Clause, and LGPL) together comprise about 39%. The remaining OSI-approved licenses constitute less than 1% of OSS licenses on Github.

MIT, GPL-3.0 and Apache have been increasing in proportion. These licenses are generally more *permissive.* Over the same period of time, GPL-2.0 and BSD, two more *restrictive* licenses, decreased in proportion.





We can only identify ~2% of the users' sectors

- The self-reported company field of the GitHub profile contained keywords we used to place each user into the above sectors.
- Of all the users that we identified, 78% of them are from the business sector, 16% are from universities.
- We can only identify about 2% of the users, suggesting that GitHub does not give enough information to accurately find the sectors for OSS contributors because users are not required to accurately fill in organization information.

Conclusions

1. **OSS is growing rapidly and becoming more permissive.**
2. **Most OSS repositories have very few contributors.**
3. **Even with perfect information on GitHub, sectors will be difficult to identify.**
4. **We need better standards for tracking and recognizing OSS producers.**

Next Steps

1. **Get more detailed data on the OSS repositories, including additions and deletions to find out the lines of code**
2. **Use the lines of code to get cost estimates of OSS Code**
3. **Encourage better standards for OSS**