MEASURING THE UNIVERSE OF OPEN SOURCE SOFTWARE

DSPG Team: Cong Cong (Illinois), Calvin Isch (Indiana University), and Eliza Tobin (University of Virginia)

SDAD Team: Gizem Korkmaz, Bayoán Santiago Calderón, Brandon Kramer, and Aaron Schroeder

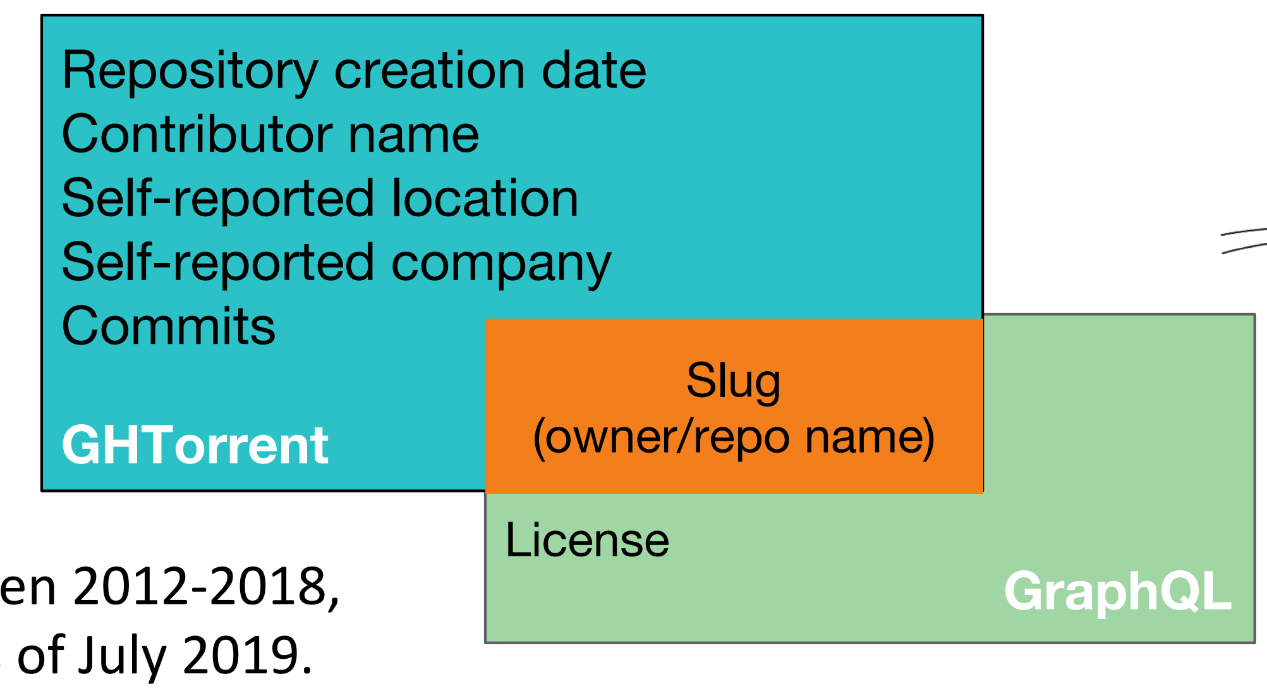
Sponsor: Carol Robbins, NCSES

Project Introduction

- This project aims to measure how much Open Source Software (OSS) is created and to better understand the distribution of OSS creation across various sectors to evaluate the economic impact of OSS.
- Traditional measures of innovation (copyrights, patents, and trademarks) do not accurately capture the universe of OSS innovation.
- We define the OSS universe as all GitHub repositories with a registered OSI-approved license.

Data Collection

- We used multiple sources to access information on GitHub repositories:
  - *GHTorrent* – an online, up-to-date database of GitHub initially funded by TUDelft & Microsoft [2]
  - *GraphQL* – GitHubs' current API system includes repository names, owners, and license information [3]



Repository creation date
Contributor name
Self-reported location
Self-reported company
Commits
GHTorrent

Slug
(owner/repo name)

License
GraphQL

We collected
- **55.1M** repositories with commits from GHTorrent between 2012-2018,
- **7M** repositories with OSI-approved licenses on GitHub as of July 2019.
  Of those, we analyzed **4.9M** repos that have at least one commit.

Terminology

Open Source Initiative (OSI) – a worldwide non-profit that spreads knowledge about OSS, promotes its usage, and connects various OSS communities [1].
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 – define limitations of use and provide developers with rights over their work while promoting the dispersion of free, accessible code. OSS licenses allow software to be freely used, modified and shared.

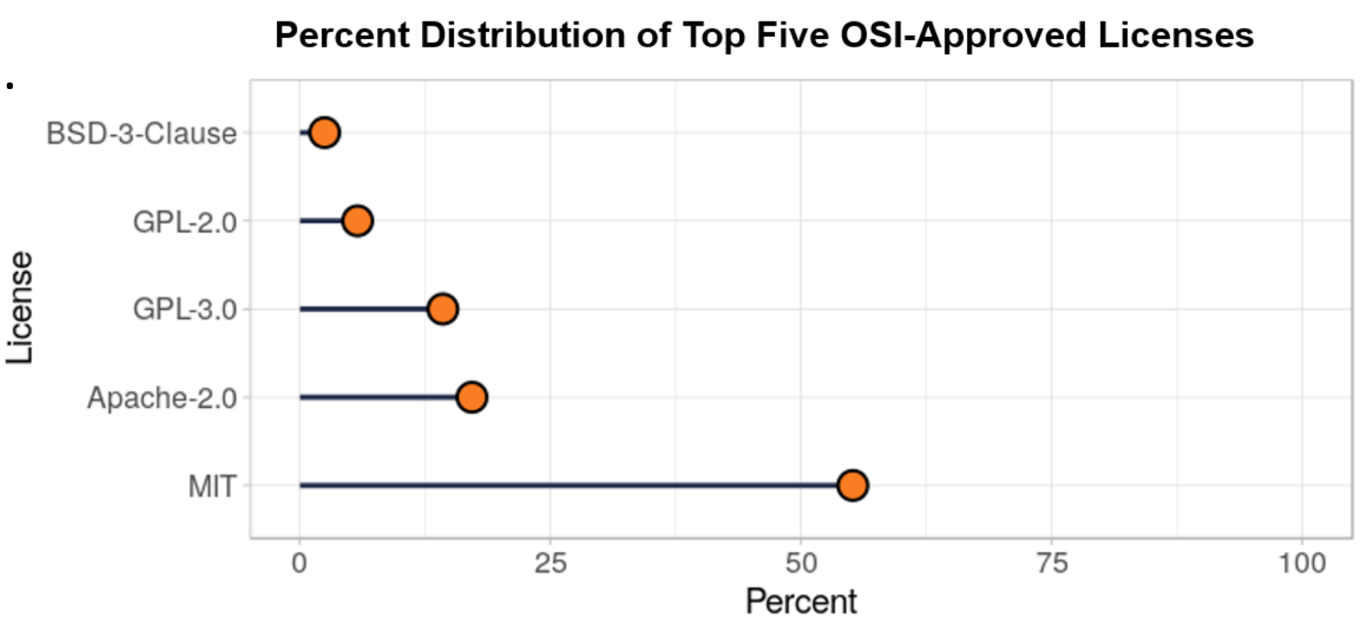
open source initiative Approved License®

GitHub - world's largest website for developers to build and share software (github.com)
Repository (Repo) - contains all of a project's files and relevant discussions
Commit - an individual change to a file for a repository
Contributor - someone who successfully committed to a project
Slug - URL friendly combination of owner and repo name (*i.e., Nosferican/Econometrics.jl)*

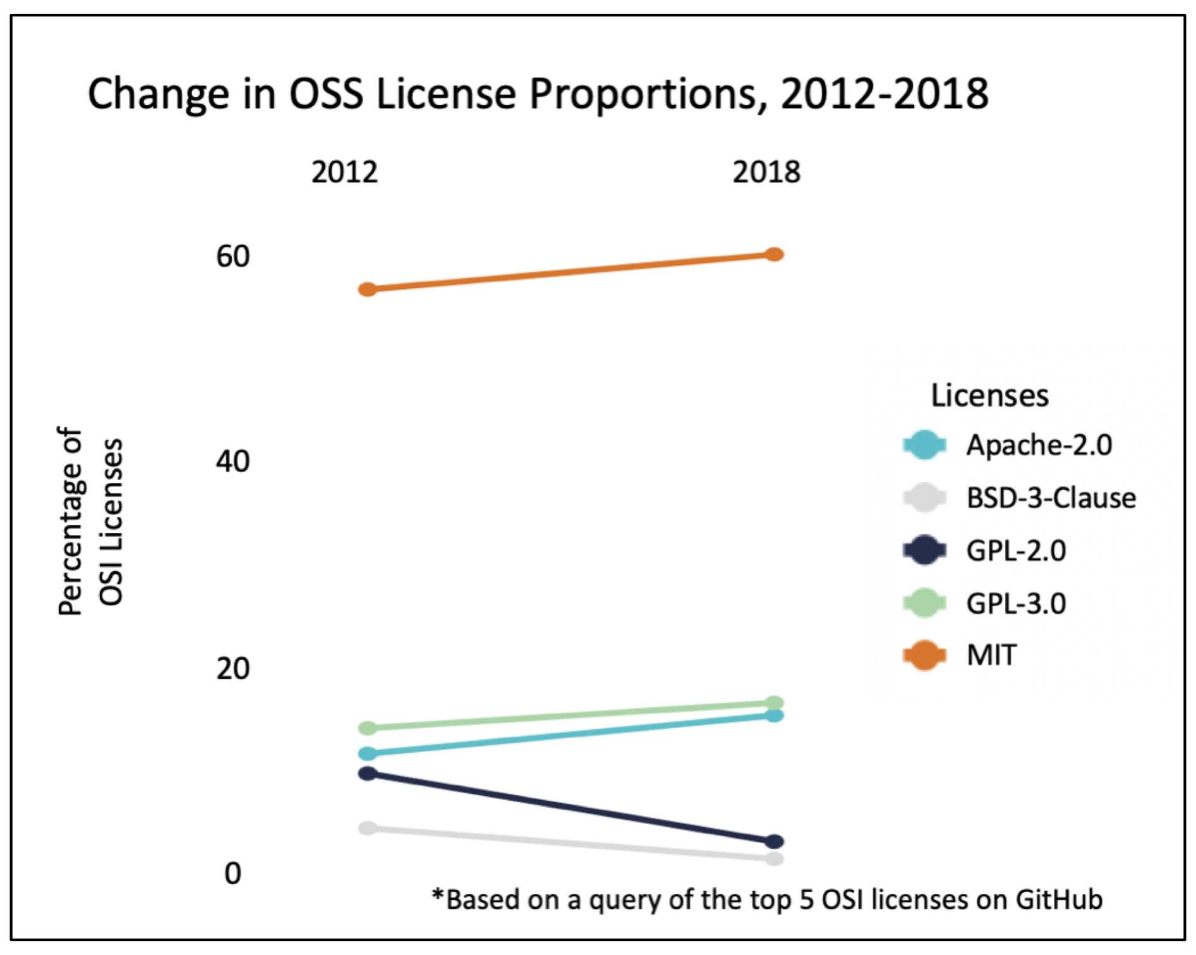
Data Analysis

Licenses

- Although 87 OSI-approved licenses exist, the top 13 licenses contain >99% of OSS repos on GitHub. The most popular ones are:
  - MIT: Massachusetts Institute of Technology
  - Apache License
  - GPL: GNU Public License
  - BSD: Berkeley Software Distribution
- License regulations vary:
  - The MIT license allows developers to use the code for any purpose.
  - The GPL license grants the ability to use the respective code under the stipulation that derivative work remains open source.



MIT is the most common OSS license (55%). These five licenses (presented above) together comprise about 93% of all OSS on GitHub.



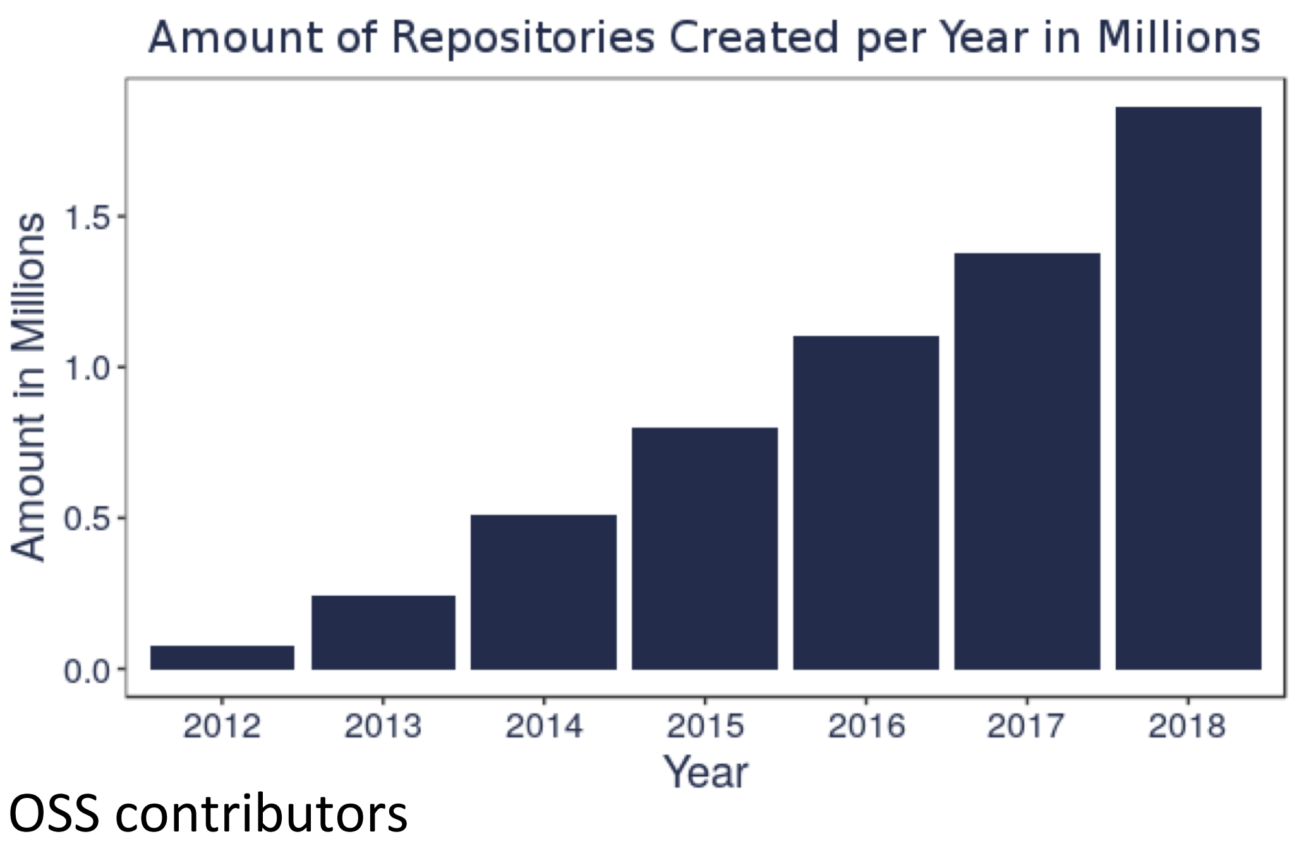
*Based on a query of the top 5 OSI licenses on GitHub

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

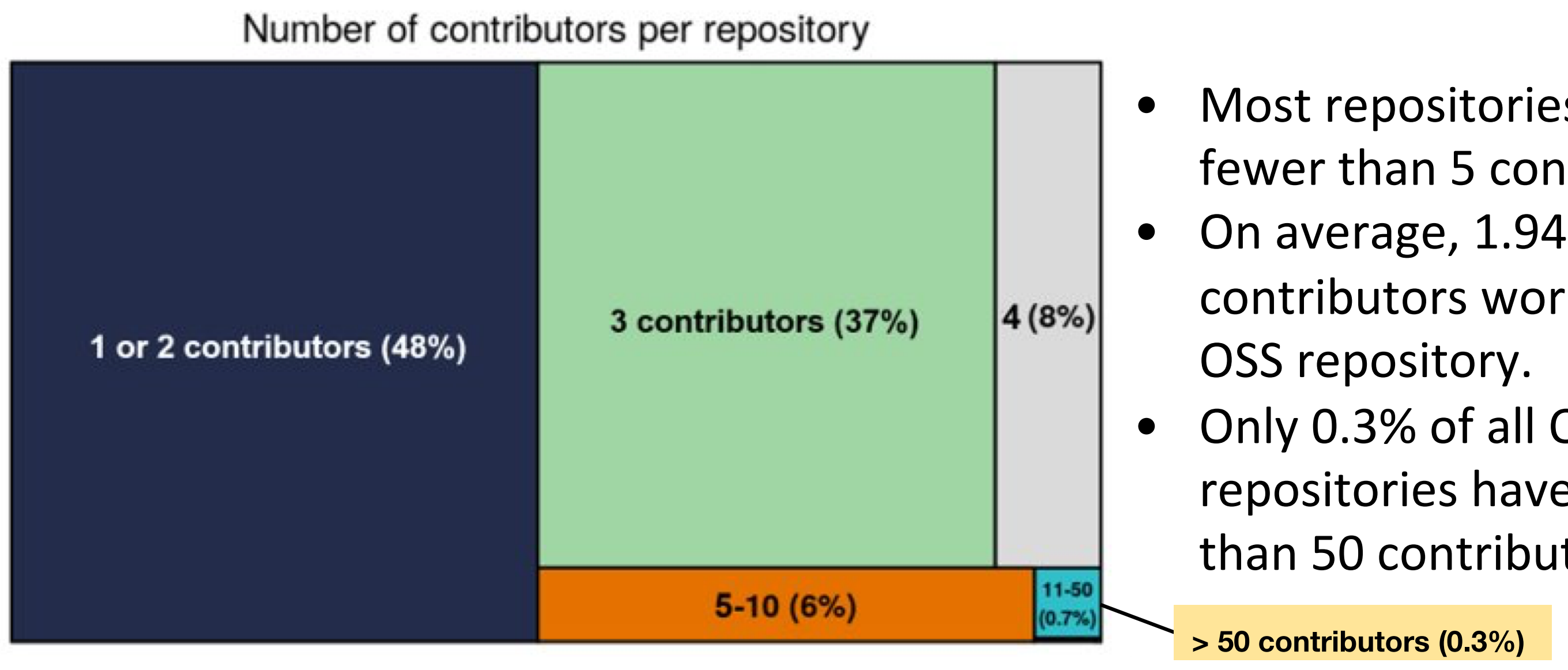
OSS Projects and Contributors

The number of GitHub repos with an OSS license has been increasing rapidly.

In 2012 there were 79.4K repos with OSS licenses. By 2018, there were 1.9M.



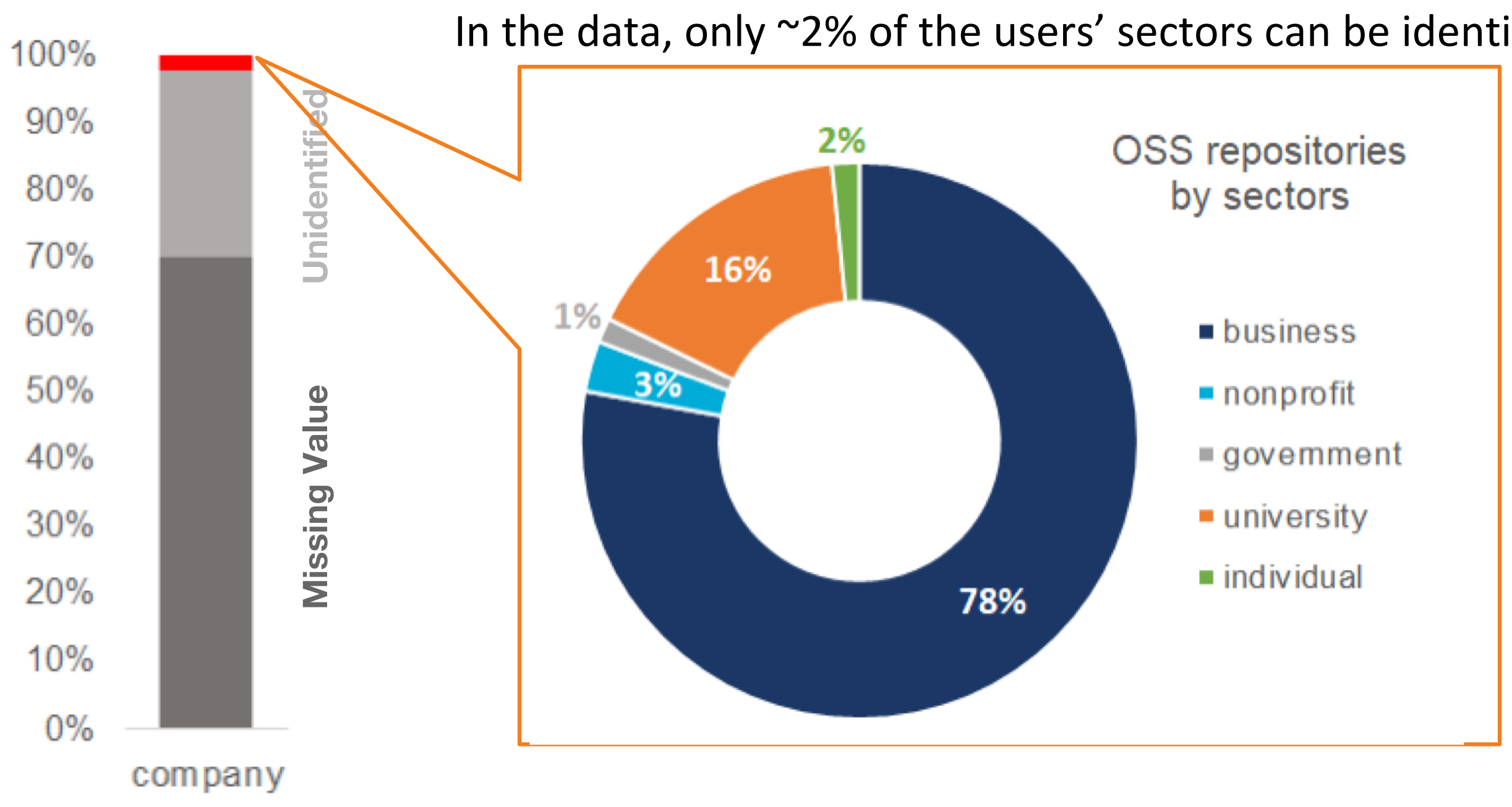
- There are 2.8M unique OSS contributors
- On average one repository receives 36.5 commits



- Most repositories have fewer than 5 contributors.
- On average, 1.94 contributors work on each OSS repository.
- Only 0.3% of all OSS repositories have more than 50 contributors.

Limitations: Sectors

In the data, only ~2% of the users' sectors can be identified



- We use the self-reported company field in contributors profile.
- Only 2% of the contributors can be identified.
- OSS is becoming more permissive as businesses contribute more code.

Conclusions

- OSS is growing rapidly; 2,350% increase in the number of repositories from 2012 to 2018.
- Permissive licenses are becoming more common (MIT is the most popular OSS license).
- Sectors are difficult to identify because users are not required to accurately fill in organization or location information.
- Better standards are needed for tracking and recognizing OSS producers.

Next Steps

- Get more detailed data on the OSS repositories, including additions and deletions to estimate the development cost using the lines of code
- Obtain contributor emails to improve the sector analysis
- Conduct network analysis to study interactions between contributors and OSS projects, and diffusion of OSS innovation

References

[1] Open Source Initiative (OSI). 1998. "The open source definition." https://opensource.org/osd.
[2] Gousios, G. 2013. "The GHTorrent dataset and tool suite." Available at http://ghtorrent.org.
[3] GraphQL. 2015. "A query language for your API." Accessed at https://developer.github.com/v4.

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE
Data Science for the Public Good