**Open Source: The Future of Software Innovation**

**2.8 Million Users Contribute to 2,350 Percent Growth over the Past 6 Years**

**Introduction**

Open source software (OSS) provides individuals with a tool to construct and modify free software while contributing to an increasingly interconnected society. Additionally, [OSS has become common practice](https://www.forbes.com/sites/forbestechcouncil/2018/07/16/how-open-source-became-the-default-business-model-for-software/#7d3bc12d4e72) in software development amongst businesses. Despite how widespread OSS is today, current economic measures do not capture the scope of OSS. As OSS becomes globally pervasive, it is vital for policymakers to measure its size and value.

Federal agencies, such as the National Center for Science and Engineering Statistics, are working to establish OSS measurement strategies, but there are many attributes of OSS that policymakers must consider. First, it is important to note that contributors come from various sectors, including businesses, universities, non-profits, governments, and individuals. Moreover, traditional methods of measuring innovation, namely copyright data, do not adequately portray the size or value of OSS, as most developers do not file copyrights for open source code. Because of the lack of adequate OSS evaluation standards, potentially highly productive OSS developers across various professions [lack project funding](https://www.nature.com/articles/d41586-019-02046-0).

In this paper, we explore the following questions:

1. How large is the OSS universe?

2. How does the universe change annually?

3. Who are the creators and where do they come from?

We define the universe of OSS as all [GitHub](https://github.com/) repositories with a registered OSI approved license. All OSS has an attached Open Source Initiative (OSI) license, and [m](https://www.wired.com/2015/03/github-conquered-google-microsoft-everyone-else/)ost open source developers store their code on GitHub, rendering it the best potential source for beginning to understand the OSS universe through the lens of a variety of licenses.

**OSS TERMINOLOGY**

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Open Source Software (OSS) | “A computer software, with its source code made available with a license” (Source: Robbins et al. 2019). |
| [Open Source Initiative](https://opensource.org/) | A worldwide non-profit that spreads knowledge about OSS, promotes its usage, and connects various OSS communities. |
| OSS License | Protections defining the limitations of use for OSS code. |
| [Repository](https://help.github.com/en/articles/github-glossary) | 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 |

**Data Collection Methods:**

There are several methods available to collect data from GitHub. We aimed to collect specific information on GitHub users and repositories that directly address our questions:

1. *Repository names*, *license, additions*, and *deletions* tell us how large the OSS Universe is.

2. With *creation date*, we can see how many repos were created over time.

3. Contributor information including *email*, *location*, and *organization* tell us who is creating OSS and what sectors they belong too.

In order to collect these data, we investigated five strategies:

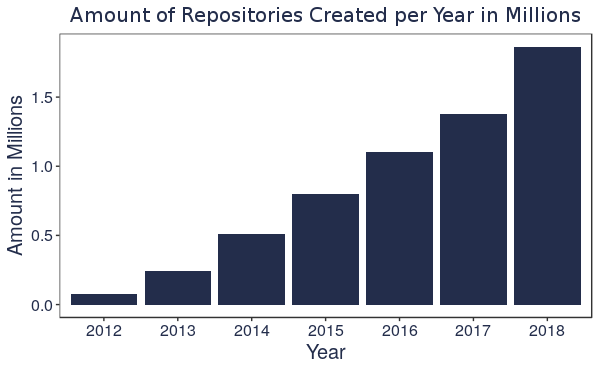
|  |  |  |  |
| --- | --- | --- | --- |
| **Method and Description** | **Information available** | **Accessibility** | **Limitations** |
| [GitHub’s **RestAPI**](https://developer.github.com/v3/) **–** a traditional API for GitHub’s data | Repository name, contributor, commits, and many more variables associated with a page of interest | Query Rest API using RStudio | -Low (5000) hourly rate limit  -Data requires more manipulation before being useful  -Cannot specify  exactly what you want (licenses) |
| GitHub’s current API: [**GraphQL**](https://developer.github.com/v4/) | Any information publicly available on the website | Query GraphQL using RStudio;  Experiment on the [explorer](https://developer.github.com/v4/explorer/). | -Limited to first 1000 results  -Low (5000 hourly rate limit)  -Frequently times out |
| Web-Scraping –crawling through HTML | Using repository names and owners, you can find the top 100 contributors and their additions and deletions | RStudio | -Time consuming  -Requires knowing specific repositories and owners |
| BigQuery – a public dataset maintained by Google | Contains all the variables accessible from RestAPI in an easy to access format | [Query the BigQuery](https://codelabs.developers.google.com/codelabs/bigquery-github/index.html?index=..%2F..index#0) using Google | -Missing data from recent years  -Data is in inaccessible RestAPI format |
| [GHTorrent](http://ghtorrent.org/) - an Open Source project with accessible GitHub information | Contains most variables accessible from a rest API | Directly downloadable datasets | -Missing license information  -Maintained by an unverified owner |

**Method**

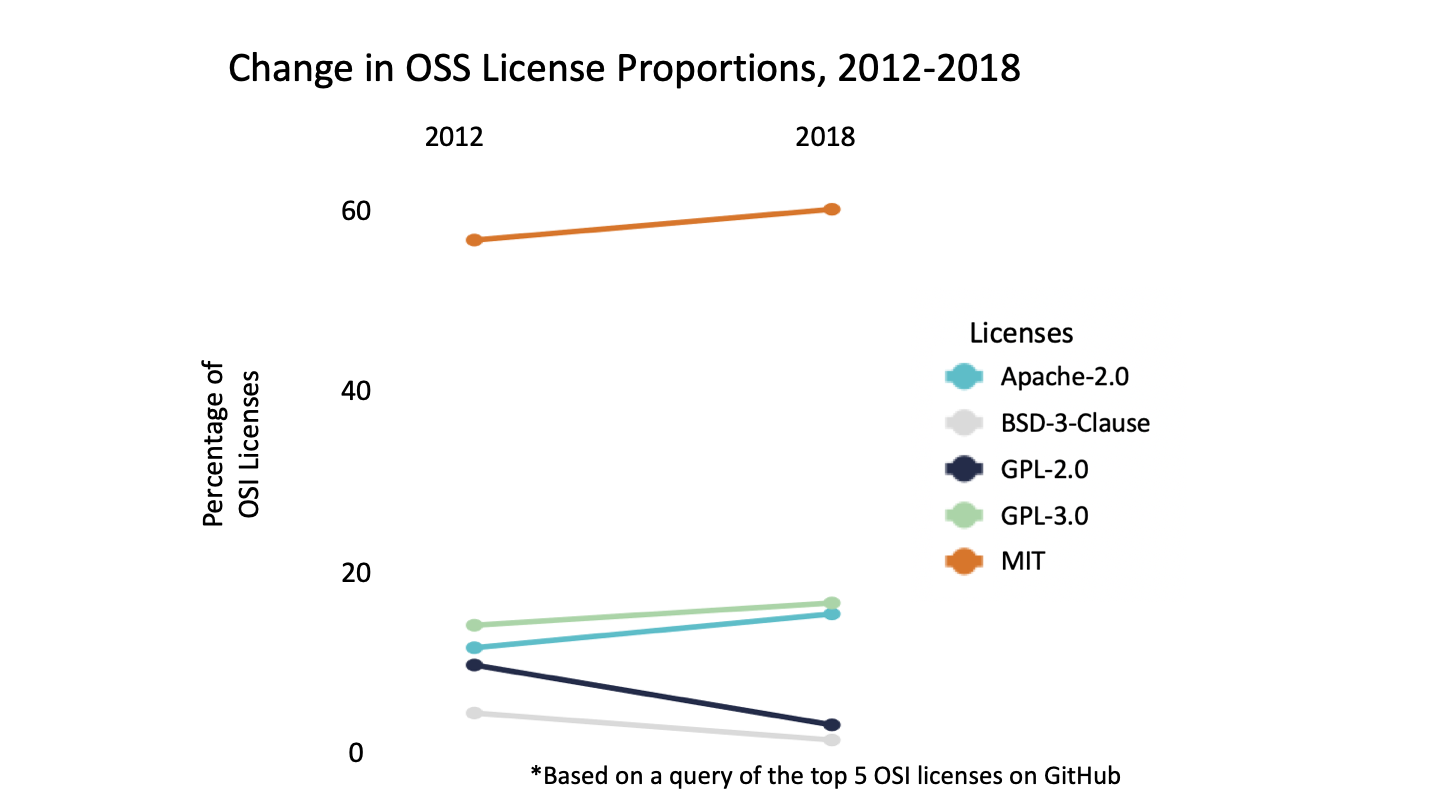
We narrowed our data collection strategy down to a combination of GraphQL and GHTorrent for three reasons: First, we ruled out the REST API because we preferred to directly search for specific licenses. We then chose the existing GHTorrent data over querying for all the information we needed to circumvent the GraphQL API issues of timeout issues and data restrictions. Finally, as the GHTorrent data do not contain OSS license information, we conducted a GraphQL query for the repository name, owner, and license. We merged the results with the GHTorrent data to obtain all the variables of interest: contributor’s location and organization, number of commits, year of repository creation, and license. This approach provides us with a robust picture of the desired information within a limited timeframe.

**Findings**

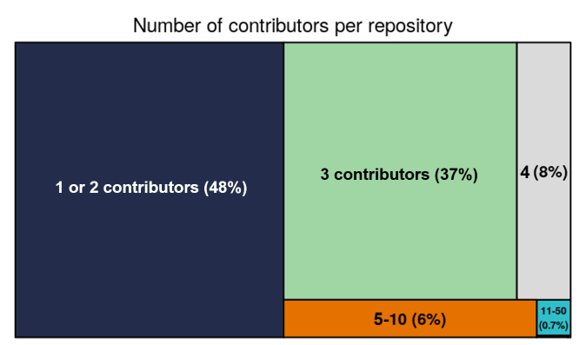
The number of OSS repositories created annually grew from 79,400 repositories made in 2012 to 1.86 million in 2018, a 2,350 percent increase (Figure 1).



The MIT license is the most commonly used license, making up about 60 percent of GitHub OSS repositories. Additionally, our findings show that more permissive licenses are growing in popularity, while more restrictive licenses are decreasing (Figure 2).



There are 2.8 million unique OSS contributors, and most repositories have fewer than five. Only 0.3 percent of all repositories have more than 50 contributors (Figure 3). This finding suggests that not only companies, but also millions of individuals, are freely producing OSS.



**Research Challenges:**

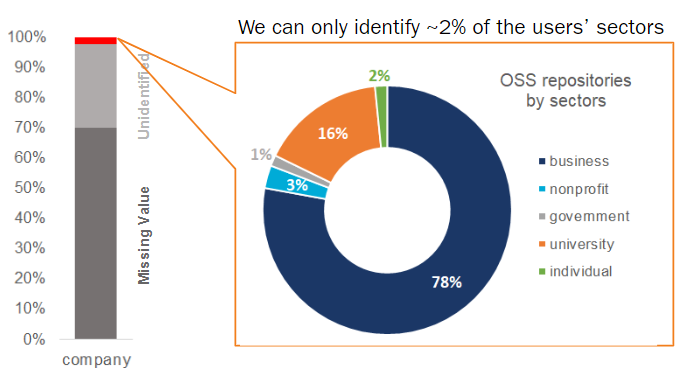
Although we acquired useful insights about the growth of OSS, data source limitations prevented us from obtaining an estimate of its economic impact, accurate measures of the sector and location for the contributors, and definitive license counts:

All the data collection methods we tried lacked data on repository’s additions or deletions or made the process technically tedious and overly time-consuming. With limited information on the lines of code to estimate developer time, we were unable to estimate the economic weight of OSS.

Additionally, although each OSI-approved license has an official name, it is possible that users list their chosen license incorrectly or using unofficial language. Thus, our counts per license are a lower-bound of the actual amount on GitHub.

Because GitHub does not require users to accurately fill out location and organization information, accurate sector identification was incredibly difficult. Figure 4 shows our sector analysis of less than two percent of organizations that we were able to identify.

Figure 4: OSS Contributor Sectors



**Next Steps and Recommendations**

Open source software is becoming more common, accessible, and globally utilized. With its explosion in size, it has become increasingly vital to measure its impact. General recording standards are lacking; however, some businesses and government entities have established tools for evaluating OSS production, providing an example to other sectors of productive measurement strategies.

Through annual reports, businesses readily report OSS output. Additionally, the U.S. government has developed [a platform](https://code.gov/) that provides an accessible record of the federal government’s software. As of July 2018, this platform housed over 4,000 unique open software projects, with an estimated development price of more than a billion dollars (Robbins et al. 2019). Still, without comprehensive databases from other sectors, researchers and government entities cannot fully grasp the total value of OSS. Sector databases should become common practice, as both contributors and research entities will benefit from having an encompassing record of OSS creation.