"A First Look at Open-Source Software Investment in the United States and in Other Countries, 2009-2019"

Carol Robbins

(National Science Foundation)

Gizem Korkmaz

(University of Virginia)

Ledia Guci

(National Science Foundation)

José B. Santiago Calderón

(University of Virginia)

Brandon L. Kramer

(University of Virginia)

Paper prepared for the IARIW-ESCoE Conference
November 11-12, 2021
Session 4
Time: Friday, November 12, 2021 [9:30-11:00 GMT+1]

A First Look at Open-Source Software Investment in the United States and in Other Countries, 2009-2019

Carol Robbins[1*], Gizem Korkmaz[2], Ledia Guci[1], J. Bayoán Santiago Calderón[3], and Brandon Kramer[2]

[1] National Center for Science and Engineering Statistics, National Science Foundation
[2] Biocomplexity Institute, University of Virginia
[3] Bureau of Economic Analysis
Disclaimer: The views expressed in this paper are those of the authors and do not necessarily reflect the position of their respective institutions.

Abstract

Development of open-source software (OSS) is widespread, and OSS is important to measure as both capital output and as an input to production. In this paper, we develop time series estimates of annual nominal and real investment and real capital stocks in OSS in the United States. We use data on OSS projects from GitHub, the largest OSS platform with over 31 million users and developers worldwide. We collect 5.2 million project repositories, containing metadata such as author, license, commits (approved code edits), and lines of code. Following methods used in software engineering to estimate the resource cost associated with creating OSS, we use lines of code and project complexity as the measure of effort to estimate the time spent on software development. In addition, we use existing methodologies for measurement of software in the U.S. national economic accounts to estimate both the annual investment and the capital stocks in software that is shared on GitHub. Our estimates show that U.S. investment in OSS was $38 billion and the real capital stock nearly $120 billion (constant 2012 U.S. dollars) in 2019.

Keywords: open-source software, federal statistics, innovation, GitHub, COCOMO

Introduction

Many OSS projects create long-lived tools that are often outputs of public spending. These are freely share-able intangible assets that in many cases have been developed outside the business sector and subsequently used within the business sector. The scale and use of these modifiable software tools highlight an aspect of technology diffusion and flow that is not captured in market measures.

As defined by the Open Source Initiative, OSS is computer software with its source code shared with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose. OSS is developed, maintained, and extended both within and outside of the private sector, through the contribution of independent developers as well as developers from universities, government research institutions, businesses, and nonprofits. Many OSS projects are developed and maintained in free repositories platforms, such as GitHub, and information embedded in these repositories, including the code, contributors, and development activity, is publicly available.

We use freely available non-survey data to estimate the scope and value of OSS through the projects hosted and shared on GitHub, the most popular source-code hosting platform with more than 31 million users and developers worldwide. The resource cost associated with creating these projects is estimated following Boehm (Boehm 1984; Boehm et al. 2000). Using a sum-of-costs approach, we first obtain an estimate of the development time per project per year based on the lines added to the source code. We then use data from the Occupational Employment and Wages Survey (OEWS) program by the Bureau of Labor Statistics to obtain an estimate of the gross payroll for computer programmers and software developers. Lastly, we apply an input factor that accounts for all other non-wage inputs (i.e., full labor costs, capital, etc.) to obtain a total resource cost estimate consistent with the methodology for own-account software in the U.S. national accounts.

This paper is organized as follows. First, we explain how software is measured in the national economic accounts and describe the project's motivation through the landscape of open-source software and the platforms where it is shared. We then describe our approach to data collection and preparation, and the methodology used to generate estimates of OSS investment and capital stocks. Next, we present and discuss these estimates along with a discussion of OSS in measured software investment and a discussion of contributors from academic and government sectors, two important public sectors. Lastly, we conclude by outlining future areas of work.

Background

Software investment in the national accounts consists of three types: prepackaged, custom, and own account (inhouse work). Unlike prepackaged and custom, own-account software is not purchased or sold; it is new, or significantly enhanced software created by business enterprises or government units for their own use and its value is estimated based on in-house expenditures for its creation (Parker et al. 2000). Development of custom and own-account software within firms and organizations brings forward new software tools. As a freely shared software tool, OSS can be custom software or own account.

This software investment drives a wide range of economic and value creation activity that challenges current measurement. Many digital products are used by consumers without a direct payment; similar to network television programming, their costs are supported by advertising. This kind of free content that is bundled with advertising can be understood as a barter transaction, content in exchange for being exposed to the advertising.

In the absence of a direct price, this content created in the business sector can be valued based on its production cost (Nakamura and Soloveichik 2015; Nakamura et al. 2017). Software and databases can provide revenue in an additional way. In use, online platforms collect data about users as well as transaction fees. These data are part of the value that the platform provides. Li and co-authors (Li et al. 2019) describe several different types of online platforms, including E-commerce, online resource sharing, e-financial services, and online social network services, where data collection provides high value to the business. In these cases, the cost and market-based approaches underestimate the value of data. Using an income approach, they argue, better captures the variety of ways that firms monetize software and data.

Beyond these categories, Corrado et al. (2005) provide a framework for consistent accounting for a larger set of intangibles that generate future benefits, including brand equity and investments in human and organizational capital. Further arguing that public expenditures yielding long-lived returns should be understood as investment, Corrado et al. (2017) propose a public investment category: information, scientific, and cultural assets. They argue that better accounting of public investment in intangibles would provide a more complete picture of economic growth.

This paper contributes to the literature on the measurement of investment in intangibles and on nonmarket digital products within the framework of the existing treatment of intellectual property products in national accounts. The estimates presented in this paper extend previous work on the resource cost of developing packages for four open-source software languages: R, Python, Julia, and JavaScript. The preliminary estimates show that the resource cost for developing packages for these languages exceeded $3 billion dollars in 2017, based on 2017 costs (Robbins et al. 2019).

Data Collection and Classification of Users on GitHub

To develop a first look at an aggregate estimate of open-source software using our bottom-up method, we collected all GitHub user activity in repositories that (1) have OSI-approved licenses, (2) were created between 2008-2019, and (3) had original content (not forked, not mirrored, and not archived) using the GHOST.jl package.[1] Once the user activity data was

[1] GHOST.il is a software tool for the Julia programming language that allows the user to collect development data for repositories using the GitHub API (https://github.com/uva-bi-sdad/Ghost.jl).

collected, we procured user data from GHTorrent (Gousios 2013), which was scraped in June 2019, and then joined those users back to all users that were in our user activity dataset.[2]

In our initial work, we used a classification process for countries and economic sectors that used string matching with regular expressions to probabilistically match self-reported user information into country codes. For the data shown in this report, our team created two packages that facilitate the classification of GitHub users (Kramer 2021). The diverstidy package provides the capacity to detect and standardize users into different geographies while the tidyorgs package detects and standardizes organizations in the academic, business, government and nonprofit sectors. Both packages rely on a "funneling" strategy to matching messy text data that speeds up the process of classifying without adding significantly more computational burden. In addition to the advantage of having public-facing software that can be shared for others to reproduce our analyses, the package reduced the lines of code that a user needs to reproduce these analyses from over 1,500 to just 5. Moreover, we observed notable increases in the overall classification performance. For example, we now classify around 78.4% of all users with any valid location, email, or company data into countries, compared to 25% using the earlier matching method. We find more than 1.2 million different contributors from more than 200 different countries or economies. Figure 1 shows the country-level distribution of the top 15 countries by number of contributors to OSS projects on GitHub between 2009 and 2019. The United States has the highest number of contributors (315 thousand), followed by China (98 thousand), India (72 thousand) and Germany (70 thousand).

[2] The GHTorrent data is named gh.ctrs_extra and the August 2021 data is named gh.ctrs_clean_0821 on the database.

Measuring Open-Source Software Investment and Capital

Open-source software cannot be measured based on observable prices and quantities or based on total revenue. The method for own-account investment in the national accounts inspires our approach. While the BEA methodology to estimate own-account software makes assumptions about the number of employees in certain occupations and industries to arrive at a time-use allocation towards software investment, our approach relies on observing the output and using the observable characteristics of the product (i.e., lines of code added during development) to come up with an estimated time-use allocation. To estimate the annual quantity of open-source software shared on GitHub, we use annual additions to the lines of code in each repository. These lines of code are translated into estimates of the person-months that would be needed to create it, based on a cost model from software engineering.

Constructive Cost Model (COCOMO)

The challenge of keeping large software projects on schedule and within budget motivates a literature in cost estimation within software engineering (Sharma et al. 2011). While costs can be estimated as a function of the number of instructions, as software projects grow, effort increases nonlinearly. Different cost models account for complexity, reliability, and scale in a variety of ways based on characteristics of the product, the platform, the contributors, and the project. Examples of these estimation models include Constructive Cost Model (COCOMO II), the Putnam Software Life Cycle Management model, and models based on function points (Boehm and Valerdi 2008). The constructive cost model is the approach that we use here. The logic of the constructive cost model is that:

Production time in Person Months = (*Calibration factor)* (*lines of code)* (*effort multipliers)*

The *calibration factor* represents the person months needed for a set number of lines of code, unadjusted for effort factors. The *effort multipliers* account for complexity, reliability, and scale for these models; they lead to increased cost.

In our use of this model, we multiply annual additions to lines of code by a COCOMO II calibration factor (Boehm et al. 2000) to estimate person months per package or project. The effort multipliers from COCOMO II are parameters chosen for the organic software class which consists of software dealing with a well-known programming language and a small, but experienced team of contributors. For the estimates presented in this paper we hold these consistent across all projects, although the model allows for these parameters to be adjusted based on additional data.[3]

The model is set up as follows:

$$\text{Effort} = 2.4\ (\text{KLOC})^{1.05}$$

$$\text{Nominal\_development\_time} = 2.5(\text{Effort})^{0.38}$$

$$\text{Development\_cost} = 2.02\ (\text{Monthly\_wage})\ (\text{Nominal\_development\_time})$$

where *KLOC* stands for kilo (thousand) lines of code. We use the number of lines added to each project as the measure of effort to estimate the nominal development time in person-months. For each OSS project, we estimate the development cost (i.e. the resource cost) by multiplying the nominal development time by monthly wages for programming occupations and an input factor that accounts for non-wage costs. We assume that the input time of contributors is roughly equivalent to the average salary for computer programmers (from Bureau of Labor Statistics

[3]This could include information about the type of software created, for example applications software versus system software.

(BLS) OEWS data) plus additional intermediate input and capital services costs.[4] Following the methodology used for own-account software at BEA, these additional costs are accounted for by multiplying the wages cost by a factor of 2.02 (Lee and Prunchak 2018).[5]

The tabulation of country-level contributors to open-source software shows the U.S. as the largest contributor based on person-months of effort (over 2 million in 2019), followed by China with over three-quarters of a million person-months of effort, and Germany, with over half a million person-months of effort in the same year (Table 1). We view this table as a reasonable cross-country comparison of effort, assuming similar skills of contributors. However, because both wage rates and price indexes for software vary substantially across countries, only the U.S. component of this activity is estimated as dollar-based investment and stocks.

Wage Series

The U.S. Bureau of Labor Statistics has 12 surveys or programs that provide information on pay and benefits. For choosing the right survey data for our purposes, we prioritize having salary and wages data at an occupational detail specific enough to capture the activities of interest and a relatively break-free time series. The Occupational Employment and Wage Statistics (OEWS) program produces employment and wage estimates annually for nearly 800 occupations and provides the necessary information to generate estimates for the period of interest (i.e., 2009 – 2019). However, the relevant wage series classification changes three times during our period, leading to time series breaks. For this paper, we use two detailed occupations that span our time

[4] See technical appendix on the choice of a wage series.

[5] In earlier work we used tables from the Input-Output from the national accounts to construct input ratios, using Computer Systems Design Services (NAICS 2017 541512). However, the data at the national industry level (6 digits) is only available during benchmark years (2007, 2012). Data for the industry group Computer Systems Design and Related Services (5415) is available annually. Using either level of detail and the corresponding versions of the use tables, the person-month resource cost is estimated. The main benefit of the current method it its improved correspondence to the BEA estimation method for own-account software, which uses a fixed estimate based on an average annual gross payroll to total expenses for NAICS 5415 using data from the Services Annual Survey (SAS).

series (2009-2019) computer programmers (BLS OEWS 15-1131) and software applications developers (BLS OEWS 15-1132) weighting each year's wage rate based on the number of workers in that occupation each year.

Investment and Stock Measures

The development cost formula described above yields an annual current dollar investment measure, for 2019, of $38 billion dollars (Table 2). Current dollar investment is divided by the annual BEA price index for own-account software, producing the estimate of constant dollar investment for the U.S., an estimated $39 billion dollars in 2019, shown in Table 3.[6] Our estimates show that real investment in OSS has grown on average nearly 50% each year since 2009, much faster than real investment in other types of software, albeit from a small base.

To create capital stock estimates, we assume that the 2009 investment value is equivalent to the 2009 initial stock value. Using a perpetual inventory method and an assumed depreciation rate of 30%, we estimate a capital stock value of open-source software of more than $118 billion dollars in 2019 (Table 4).

Open-Source Investment in Measured Software Investment

In economic accounts of intangible investment and intellectual property products, software investment is measured in three types, prepackaged, custom, and own account. Prepackaged and custom software are purchased inputs, and in national economic accounts, industry receipts and government budget data are used for these estimates. For software investment in 2019 BEA reports a total of $492 billion dollars, $428 of which is accounted for by private investment, a

[6] BEA NIPA Table 5.6.4. Price Indexes for Private Fixed Investment in Intellectual Property Products by Type, July 2021.

category which includes private businesses and non-profit institutions serving households (NPISHs). Prepackaged software accounts for 44% of this private investment, with custom and own-account software making up 39% and 17%, respectively.

Government investment in software reported by BEA for 2019 is $41.8 billion dollars for federal and $22.8 billion for state and local government. In the U.S. economic accounts, public university investment is counted within state and local or federal (military-affiliated colleges, for example). Our estimates from both academia and government suggest that these categories may be underestimated.

Based on our estimates, open-source software is growing more rapidly than the NIPA measures of prepackaged, custom, and own-account software, suggesting that if these current trends continue open-source software will become an increasing source of software investment in the future (Figure 3).

An unavoidable challenge in estimating own-account production of software is the potential for overlap between own-account software and own-account or inhouse research and development (R&D) activity. This overlap is estimated with business R&D survey data, among other sources. This overlap for open-source software, along with overlap between business software in the national accounts and the open-source software we measure based on GitHub will offset some the impact to $38 billion dollars in investment in open-source software (in 2019) that we estimate in this paper.

Academic Contributors

For many academics and researchers, software tools and databases are by-products of their own work that can also be used by other academics as well (Gambardella and Hall 2005). When these research tools are shared freely, they represent a free input from the perspective of the software user. Between 2009 and 2019 almost 45,000 individuals with U.S. academic affiliations contributed to OSS repositories on GitHub (Figure 4). The country with the next largest number of contributors is China, with more than 10,000, followed by the United Kingdom with more than 5,000 academic contributors.

By individual institution, the highest number of academic contributors are affiliated with the University of California at Berkeley, followed by MIT, Carnegie Mellon and Stanford (Figure 5). Universities from outside of the United States in the top 20 contributors are two Canadian Universities, the University of Toronto and the University of Waterloo, and three universities in China: Shanghai Jiaotong, Zhejiang, and Tsinghua.

Government Contributors

An additional and substantial source of open-source software tools is the U.S. federal government, which supports the sharing of software developed by and for the federal government through its Federal Source Code Policy. This policy provides a framework for government code to be released and reused through open-source software (OSS) licensing, allowing software created for narrow federal purposes to be reused elsewhere within the federal government, multiplying its value to the government, and outside of the federal government, further extending its impact. When this software is used outside of the federal government, again, it represents a free input from the perspective of the software user.

Units of the U.S. Federal government share software through well-known platforms, such as GitHub, SourceForge, and Bitbucket, as well as webpage repositories run by units of the federal government, such as those run by the National Aeronautics and Space Administration (NASA) and Sandia National Lab. Only part of the federal contributions, those that are available on

GitHub (currently the world's largest OSS hosting platform) are included in the estimates in this paper, implying an incomplete count.

Consistent with the methods described for academic contributors, repository-level data were accessed through the GHOST.jl (Santiago Calderón 2021) software package. A concordance file was prepared using *The United States Government Manual*, the A-Z Index of U.S. Government Departments and Agencies (usa.gov), the GitHub crowd-source government entities directory, and code.gov, the Federal government website that lists federal open-source projects.

An additional undercount to this method is that matching using the authors' emails to identify their affiliation allows the tabulated data to include contributions based on the assumption that they are contributing as part of that organization. If business-related addresses were used, as for government contractors that create shared software, these additions would not be counted.

Measured by the number of repositories to which government-affiliated individuals contribute, the U.S. Department of Energy has the largest number of open-source repositories on GitHub, a total of more than 11,000 between 2010 and 2019, followed by the National Aeronautics and Space Administration with over 1,000 (Table 6). As a measure of comparison, the largest private contributors where Microsoft and RedHat, each with about 25,000 repositories with contributions between 2010 and 2019. The University of California at Berkeley contributed to about 7,000 repositories in the same period.

Future Work

The GitHub data, which we plan to make publicly available, present a unique opportunity to conduct a wealth of analyses around the development of open-source software. We plan to focus our efforts in the following areas.

First, complete the classification of OSS contributors to sectors and estimate the contribution of each sector to OSS. Of particular interest is the contribution to OSS from public spending so an area of priority is estimating the amount of OSS shared by the U.S. federal government.

Second, refine the estimation approach by modifying the model to account for different types of software and different cost parameters. The fixed effort parameters from a "typical" software project are first approximations. Some types of software may require greater effort. Systems software, which includes operating systems, networking software, database management and development tools may require greater effort to function on multiple platforms or require more frequent upgrading than some types of applications software. Applications software includes business and home financial software, statistical software and gaming software.[7]

Third, asses the overlap of OSS with custom and own-account software as well as the overlap with R&D investment to better assess the impact that OSS has on GDP.

Finally, use OSS contributors' locations to generate contributor networks and study structural features of international collaborations using social network analysis methods and identify key players in the OSS ecosystem.

[7] One piece of evidence suggesting that systems software requires higher than average skill is that the annual wage for systems developers is higher than that of applications software developers in BLS wage data.

Summary

We estimate a stock of zero price open-source software tools shared on the GitHub repository between 2009 and 2019 at over $100 billion dollars for the U.S. Using a bottom-up estimation method based on the quantity of code and fixed effort parameters from software engineering literature, we also estimate annual labor effort contributions to open-source software on GitHub for over 200 countries or regions.

The estimation of effort and resource costs provide what we believe to be a lower boundary on the overall value of open-source software in the U.S. economy. The method we use is designed to be consistent with the existing boundaries of produced assets in national accounts, treating open-source software as an asset used by its creators, with an additional zero price benefit for those who share and use the open-source software.

Through matching of contributors, institutional affiliations, and email addresses we can assign an increasing share of the open-source projects to economic sectors, finding large numbers of projects on GitHub generated from universities and government institutions.

We view our contribution as an incremental decrease in the scope of unaccounted for intangible assets that contribute to economic growth. With additional open-source contributions as yet uncounted, we find the magnitude and growth of open-source software is big enough and important enough to more accurately account for its role as both an investment output and an unpriced capital input.

Acknowledgements

This material is based on work supported by U.S. Department of Agriculture (58-3AEU-7-0074). The authors acknowledge Research Computing at the University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication. We also acknowledge the Data Science for the Public Good Program participants.

References

Apache License, 2019. Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.

Boehm, B.W., 1984. Software engineering economics. IEEE Transactions on Software Engineering SE-10, 4–21. doi:10.1109/TSE.1984.5010193

Boehm, B.W., Clark, Horowitz, Brown, Reifer, Chulani, Madachy, R., Steece, B., 2000. Software Cost Estimation with COCOMO II (with CD-ROM). 1st ed., Prentice Hall, Upper Saddle River, NJ, USA.

Boehm, B.W., Valerdi, R., 2008. Achievements and challenges in COCOMO-based software resource estimation. IEEE software 25, 74–83. BSD Licenses, 2019. Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index. [Online; accessed 26-January-2019].

Corrado, C., Haskel, J., Jona-Lasinio, C., 2015. Public intangibles: The public sector and economic growth in the SNA, in: Economics Program Working Paper Series. The Conference Board. Available at https://www.conference-board.org/pdf_free/workingpapers/EPWP1501.pdf.

Corrado, C., Haskel, J., Jona-Lasinio, C., 2017. Public intangibles: The public sector and economic growth in the SNA. Review of Income and Wealth 63, 355–380. Corrado, C., Hulten, C., Sichel, D., 2005. Measuring capital and technology: An expanded framework,

Gambardella, A., Hall, B.H., 2006. Proprietary versus public domain licensing of software and research products. Research Policy 35, 875–892. Gault, F., 2018. Defining and measuring innovation in all sectors of the economy. Research Policy 47, 617–622.

Gousios, G. 2013. "The GH Torent Dataset and Tool Suite." Proceedings of the 10th Working Conference on Mining Software Repositories, Available at https://www.gousios.gr/pub/ghtorrent-dataset-toolsuite.pdf

GNU General Public License, 2019. Wikipedia, the free encyclopedia. https://en.wikipedia.

Kramer, B. 2021. diverstidy: A tidy package for detection and standardization of geographic, population, and diversity-related terminology in unstructured text data.
URL: https://github.com/brandonleekramer/diverstidy

Kramer, B. 2021. tidyorgs: A tidy package that standardizes text data for organizational analysis.
URL: https://github.com/brandonleekramer/tidyorgs

Lee J and Prunchak G. 2018. "2018 Comprehensive Revision - Software R&D Overlap Proposal, Memorandum to the National Accounts Methods Board." BEA memorandum.

Li, W.C., Nirei, M., Yamana, K., 2019. Value of data: There's no such thing as a free lunch in the digital economy. Discussion Paper Series 19-E-022. Research Institute of Economy, Trade and Industry (RIETI).

Martin, B.R., 2016. Twenty challenges for innovation studies. Science and Public Policy 43, 432–450.

MIT License, 2019. Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=MIT_License&oldid=881603526. [Online; accessed 26-January-2019].

Nakamura, L.I., Samuels, J., Soloveichik, R.H., 2017. Measuring the 'free' digital economy within the gdp and productivity accounts. Working Paper. Federal Reserve Bank of Philadelphia.

Octoverse, 2018. The state of the Octoverse. URL: https://octoverse.github.com. Open Source Initiative, 1998a. URL: https://opensource.org/osd. Open Source Initiative, 1998b. Licenses & standards. https://opensource.org/licenses.

Parker, R.P., Grimm, B.T., et al., 2000. Recognition of Business and Government Expenditures for Software as Investment: Methodology and Quantitative Impacts, 1959-98. Technical Report. Bureau of Economic Analysis

Robbins, C., G. Korkmaz, J. Santiago Calderón, D. Chen, A. Schroeder C. Kelling, S. Shipp, S. Keller. 2019. The Scope and Impact of Open Source Software as Intangible Capital: A Framework for Measurement with an Application Based on the Use of R Packages. NBER Conference on Research on Income and Wealth, Bethesda MD, March 15-16, 2019.

Santiago Calderón, J.B. 2021. GHOST.jl: GitHub Open-Source Software Tracker. URL: https://github.com/uva-bi-sdad/GHOST.jl.

Technical Appendix

OSI-Approved and Machine Detectable Licenses

- SPDX: Software Package Data Exchange
- 0BSD: BSD Zero Clause License
- AFL-3.0: Academic Free License v3.0
- AGPL-3.0 GNU: Affero General Public License v3.0
- Apache-2.0: Apache License 2.0
- Artistic-2.0: Artistic License 2.0
- BSD-2-Clause: BSD 2-Clause “Simplified” License
- BSD-3-Clause: BSD 3-Clause “New” or “Revised” License
- BSL-1.0: Boost Software License 1.0
- CECILL-2.1: CeCILL Free Software License Agreement v2.1
- ECL-2.0 Educational Community License v2.0
- EPL-1.0: Eclipse Public License 1.0
- EPL-2.0: Eclipse Public License 2.0
- EUPL-1.1: European Union Public License 1.1
- EUPL-1.2: European Union Public License 1.2
- GPL-2.0: GNU General Public License v2.0 only
- GPL-3.0: GNU General Public License v3.0 only
- ISC ISC: License
- LGPL-2.1: GNU Lesser General Public License v2.1 only
- LGPL-3.0: GNU Lesser General Public License v3.0 only
- LPPL-1.3c: LaTeX Project Public License v1.3c
- MIT: MIT License
- MPL-2.0: Mozilla Public License 2.0
- MS-PL: Microsoft Public License
- MS-RL: Microsoft Reciprocal License
- NCSA: University of Illinois/NCSA Open-Source License
- OFL-1.1: SIL Open Font License 1.1
- OSL-3.0: Open Software License 3.0
- PostgreSQL: PostgreSQL License

- UPL-1.0: Universal Permissive License v1.0
- Unlicense: The Unlicense
- Zlib: zlib License

Table 1. Open-Source Software Created by Selected Country, Measured in Person-months of Effort: 2009 to 2019

(Person-months)

Country	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
United States	52,133	81,701	138,360	225,320	466,468	872,968	1,200,963	1,472,416	1,812,367	2,133,874	2,135,120
China	2,352	3,310	6,096	16,918	61,041	143,679	243,433	401,661	577,745	735,792	757,258
Germany	15,813	25,055	38,262	59,659	111,251	191,967	275,398	348,045	444,566	528,541	552,139
United Kingdom	11,412	19,915	32,930	50,863	107,112	193,214	264,572	325,064	404,751	474,029	482,790
India	826	1,679	3,720	7,899	23,266	58,887	97,892	169,852	285,308	409,959	452,277
Brazil	1,392	2,993	6,201	9,480	26,181	60,146	103,444	152,922	229,063	301,346	380,377
Canada	6,812	9,948	16,732	30,607	60,994	119,561	169,995	222,021	276,609	339,405	348,002
France	6,311	11,695	19,705	32,279	67,318	127,444	173,827	212,318	266,841	313,660	319,025
Japan	3,451	5,201	9,312	17,796	35,465	71,515	100,863	125,943	163,193	212,768	232,303
Russia	2,297	4,006	6,855	14,729	33,094	57,396	83,325	124,362	168,825	202,691	229,340
Spain	2,271	3,869	7,942	14,852	31,490	61,650	94,940	122,105	154,655	188,678	192,029
Indonesia	110	163	769	1,114	4,540	11,745	27,255	48,829	87,840	133,200	184,835
Australia	4,600	7,263	11,519	18,432	36,920	64,661	88,906	110,170	139,514	166,116	176,172
Netherlands	3,929	6,601	11,423	18,341	34,253	65,657	97,914	122,789	146,623	168,132	173,490
Poland	1,442	2,487	4,916	7,775	18,208	36,128	50,345	68,269	94,243	120,563	128,894
Italy	1,924	3,404	6,407	10,030	19,937	39,576	59,489	75,786	94,048	117,856	125,773
Switzerland	3,280	5,756	9,265	15,241	27,719	46,464	62,811	77,414	104,376	117,768	116,455
South Korea	298	351	750	1,787	5,640	15,102	23,987	36,528	53,000	81,141	106,546
Sweden	2,848	4,754	7,987	11,942	24,523	46,448	61,602	73,712	86,533	104,132	105,795
Ukraine	458	1,173	2,383	3,993	11,103	38,324	32,513	47,272	68,287	82,504	92,561
Mexico	623	1,039	1,935	3,043	8,571	19,402	31,135	44,487	56,465	76,031	84,426
Colombia	608	966	1,979	3,432	8,316	18,547	28,954	40,386	59,671	79,905	80,824
Turkey	241	278	939	1,326	3,836	9,160	13,441	24,026	38,356	58,031	75,045
Argentina	767	1,374	2,955	3,917	11,142	22,017	30,002	37,660	44,464	58,480	67,606
Norway	2,145	3,271	4,507	6,505	11,875	21,721	29,616	36,256	46,076	57,930	62,909
Belgium	1,855	2,686	4,172	6,827	13,961	23,808	32,334	40,285	49,406	58,443	61,896
Czech Republic	1,553	2,554	4,634	6,053	11,380	21,301	29,197	38,558	45,841	53,753	58,357
Finland	1,745	2,149	3,715	6,680	11,809	22,049	28,117	34,252	43,772	53,247	55,891
Taiwan	560	874	1,527	2,725	7,701	18,279	36,968	37,443	49,878	57,237	54,262
Austria	1,676	2,312	4,077	5,978	12,522	20,310	26,457	34,777	43,935	50,356	52,540
Not assignable to country	277,780	357,459	456,811	637,368	1,251,594	2,385,662	3,406,534	4,928,790	6,818,480	9,001,735	9,763,550
assignable to country	427,388	599,039	865,114	1,318,932	2,708,479	5,206,389	7,471,708	10,283,537	13,871,973	17,731,093	19,009,746

Table 2. U.S. Investment in Open-Source Software, by Producing Sector: 2009 to 2019

(Millions of US dollars)

Producing Sector	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
All Sectors	733	1,158	2,005	3,338	7,165	13,870	19,649	24,645	31,234	37,472	38,440
Private											
Domestic Business											
Nonprofits serving households											
Households											
Government											
Federal											
State and local											
Addenda											
Universities and Colleges											
Private											
Public											
Note: Investment is estimated as person-months x wage rate x 2.02											
monthly resource cost in millions	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Resource cost = monthly wage x 2.02	14,068.30	14,173.23	14,493.24	14,815.26	15,360.81	15,887.84	16,360.80	16,737.73	17,233.81	17,560.35	18,003.90
Monthly wage	6,964.51	7,016.45	7,174.87	7,334.29	7,604.36	7,865.27	8,099.40	8,286.00	8,531.59	8,693.24	8,912.82
Authors' Annual Wage Series	83,574	84,197	86,098	88,011	91,252	94,383	97,193	99,432	102,379	104,319	106,954

Table 3. Real Investment in Open-Source Software in the United States, by Producing Sector: 2009 to 2019

(Millions of constant 2012 US dollars)

Producing Sector	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
All Sectors	729	1,166	2,005	3,338	7,156	13,869	19,708	24,887	31,768	38,224	39,144
Private											
Domestic Business											
Nonprofits serving households											
Households											
Government											
Federal											
State and local											
Addendum											
Universities and Colleges											
Private											
Public											
Own Account Software Price Index	100.665	99.293	100.004	100	100.127	100.002	99.701	99.029	98.319	98.031	98.202

Note: Real investment = (Nominal/Price Index) * 100

Source: BEA NIPA Table Table 5.6.4. Price Indexes for Private Fixed Investment in Intellectual Property Products by Type, version date July 30, 2021, Index for Own Account Software

Table 4. Real Capital Stock in Open-Source Software, by Producing Sector: 2009 to 2019

(Millions of constant 2012 US dollars)

Producing Sector	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
All Sectors	729	1,749	3,404	6,062	12,006	23,474	38,487	55,676	76,309	99,271	118,561
Private											
Domestic Business											
Nonprofits serving households											
Households											
Government											
Federal											
State and local											
Addenda											
Universities and Colleges											
Private											
Public											

Notes: 2009 initial value is assumed to be the 2009 real investment value.
Capital stock in subsequent years = last year's value *(1-d) + this years investment.
The assumed depreciation rate is 0.3 per year.

Table 5. Number of U.S. and Foreign GitHub Contributors by Producing Sector

Number of Contributors

Producing Sector	U.S. Contributors	Contributors
All Sectors		
Private		
Domestic Business	56,631	127,612
Nonprofits serving households	520	637
Households	1,383	7,324
Government	615	482
Addenda		
Universities and Colleges	43,512	61,562
Private	19,031	NA
Public	24,300	NA

Note: Users with multiple country or organizational affiliations are fractionally counted across the relevant countries or organizations.

Table 6. Cumulative contribution of selected entities to open-source software on GitHub: 2010–2019

(Number)

Institution	Count of repositories contributed to by institutional members
Department of Energy	11,156
National Aeronautics and Space Administration	1,102
Department of Health and Human Services	863
Department of Commerce	819
Department of the Interior	537
Department of Defense	321
General Services Administration	319
Smithsonian Institution	107
Department of Agriculture	104
Department of Veterans Affairs	76
All other federal departments and agencies	312
Federal total	15,716
Microsoft	25,365
RedHat	24,767
UC Berkeley	7,152

Note(s):
Included repositories are those public on GitHub with machine detectable Open Source Initiative-approved licenses that received contributions from affiliates from each organization; software shared through other licenses are excluded.
Public higher education institutions and selected private entities are included as benchmarks. The Department of Energy encompasses 17 national laboratories, 16 of which were included in the data. The exception is the Department of Energy's SLAC National Accelerator Laboratory, formerly known as Stanford Linear Accelerator. Due to data constraints, these repositories are attributed to the operator (Stanford), rather than to the Department of Energy. Overall, in terms of number of repositories, Microsoft and RedHat are the two largest contributors from the private sector and UC Berkeley from higher education institutions.

Figure 1.Total Contributors to OSS projects on GitHub by Top 15 Countries, 2009-2019

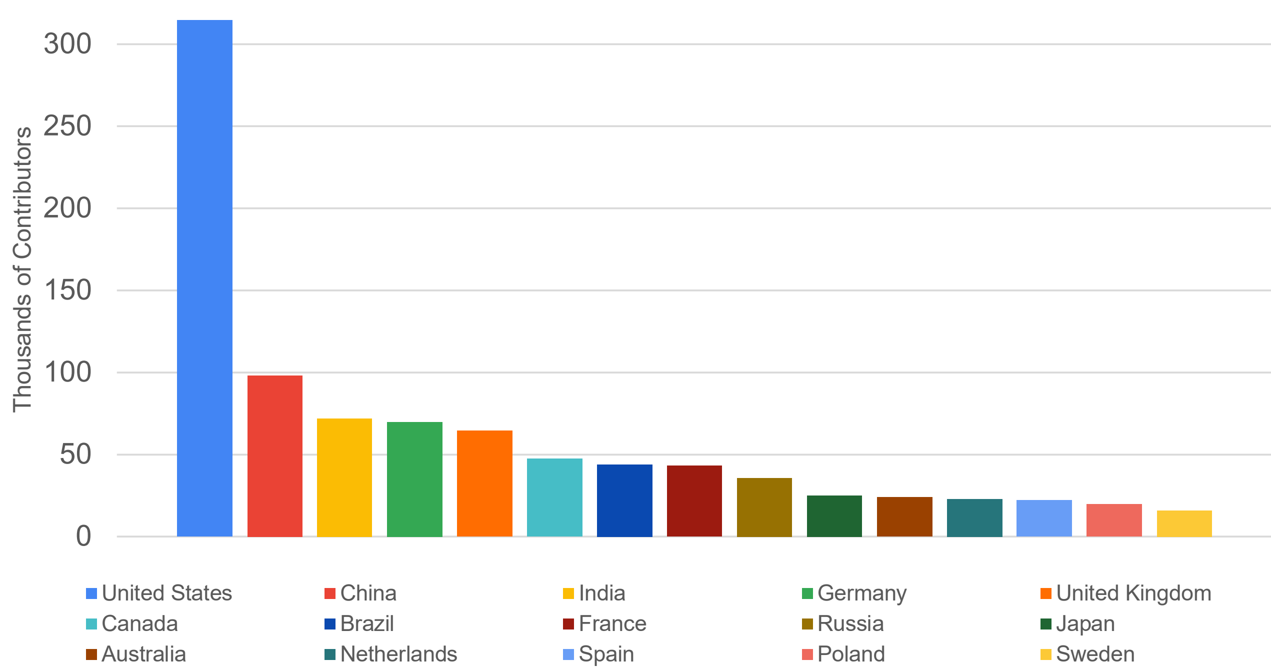


Figure 2. Top-10 Country Countributors to OSS on GitHub by person-month: 2009 -2019

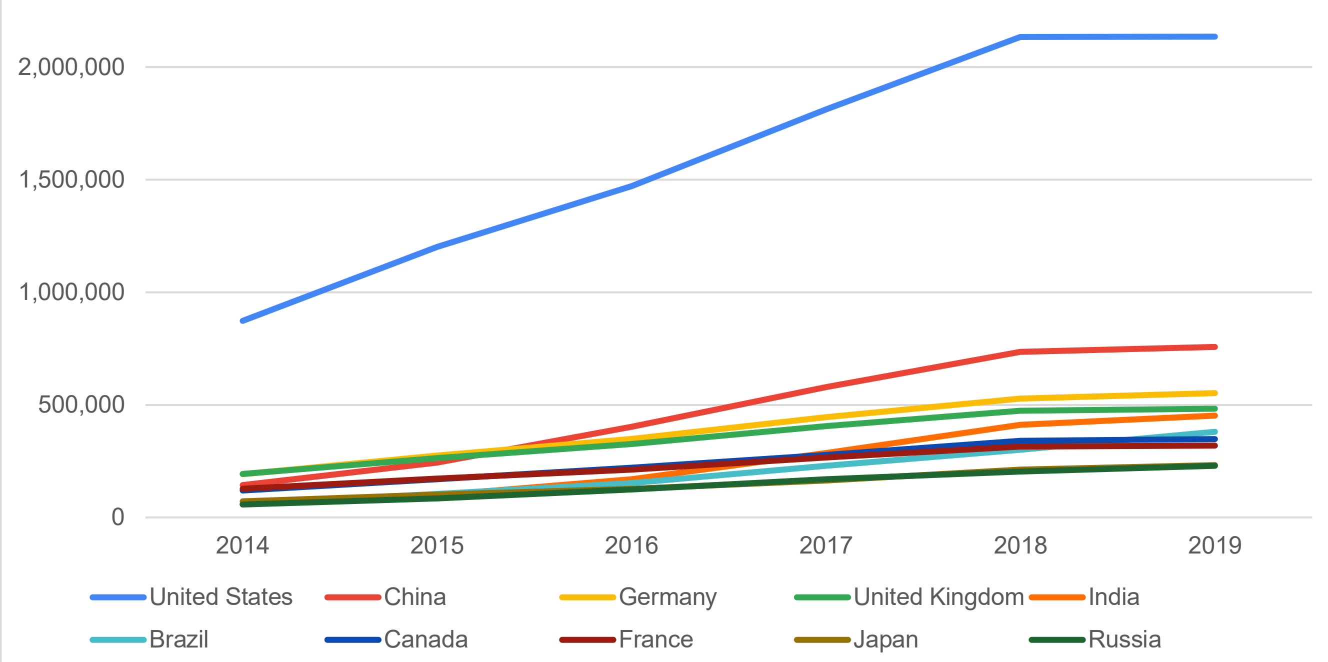
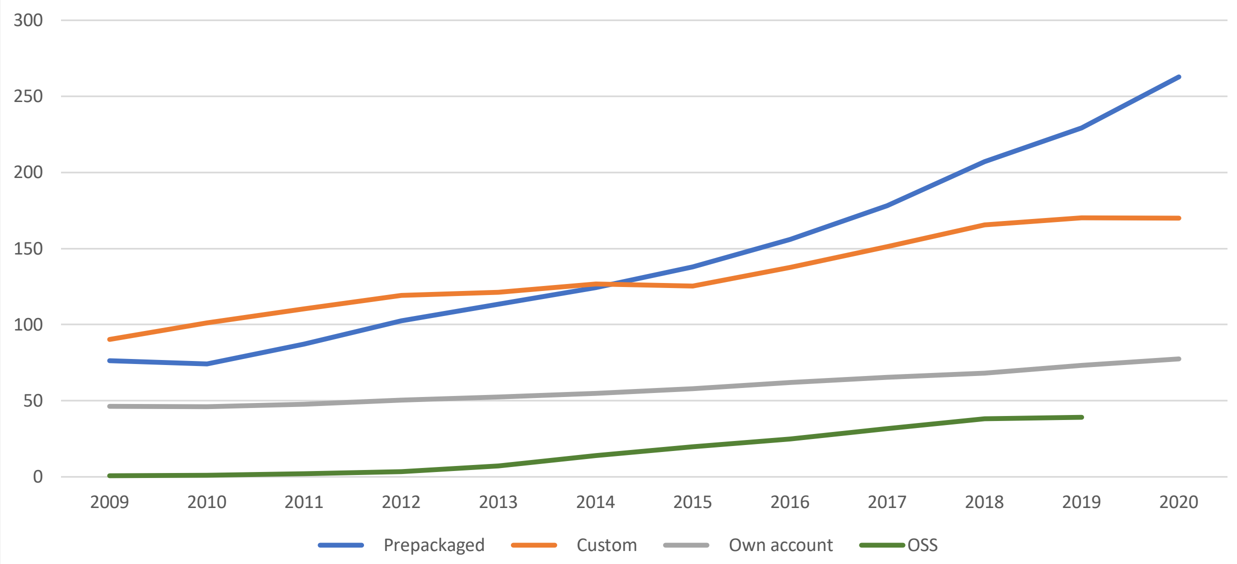


Figure 3. Real Investment in Open-Source Software and Software by NIPA type
[Billions of 2012 constant dollars]



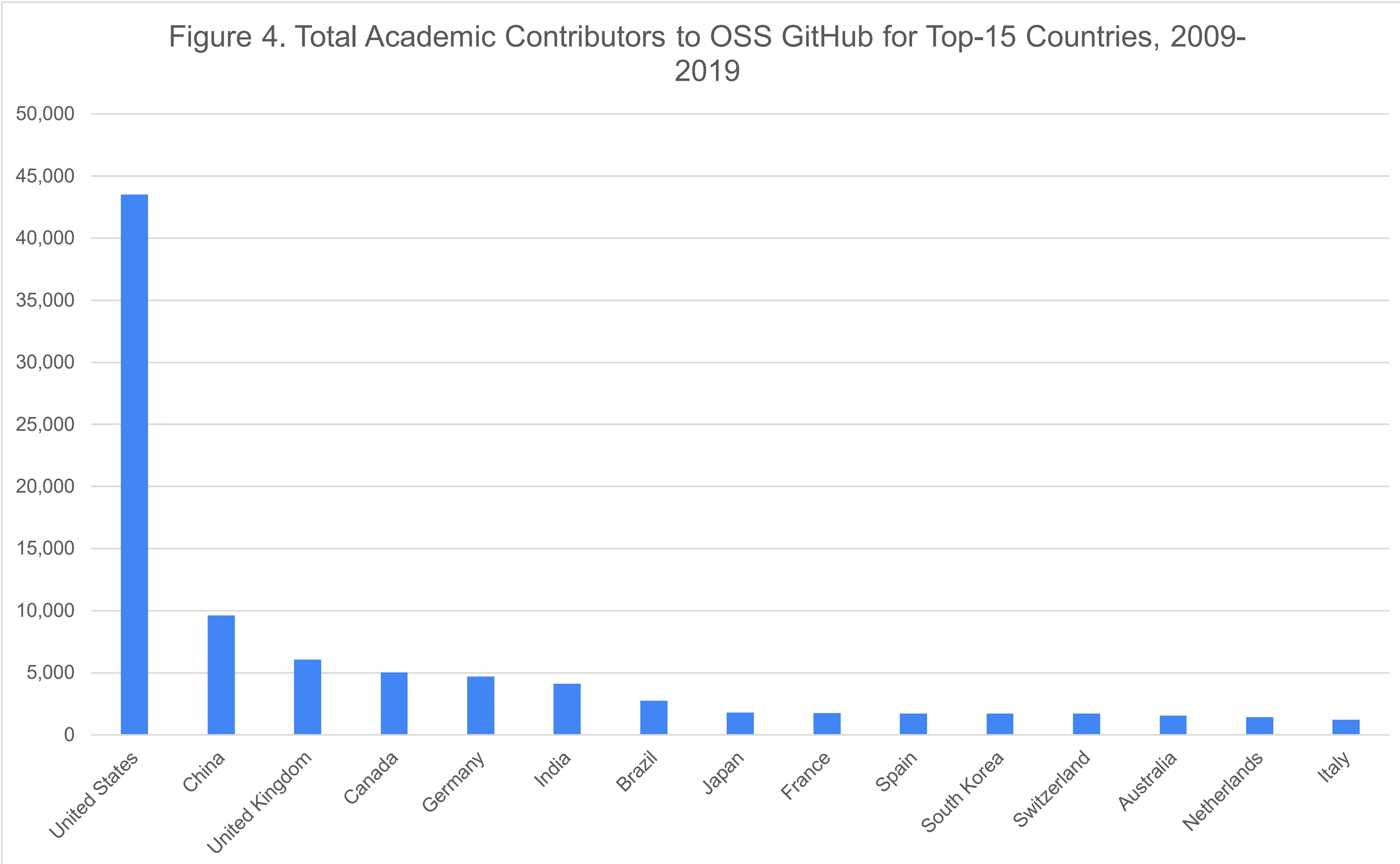


Figure 4. Total Academic Contributors to OSS GitHub for Top-15 Countries, 2009-2019
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
United States
China
United Kingdom
Canada
Germany
India
Brazil
Japan
France
Spain
South Korea
Switzerland
Australia
Netherlands
Italy

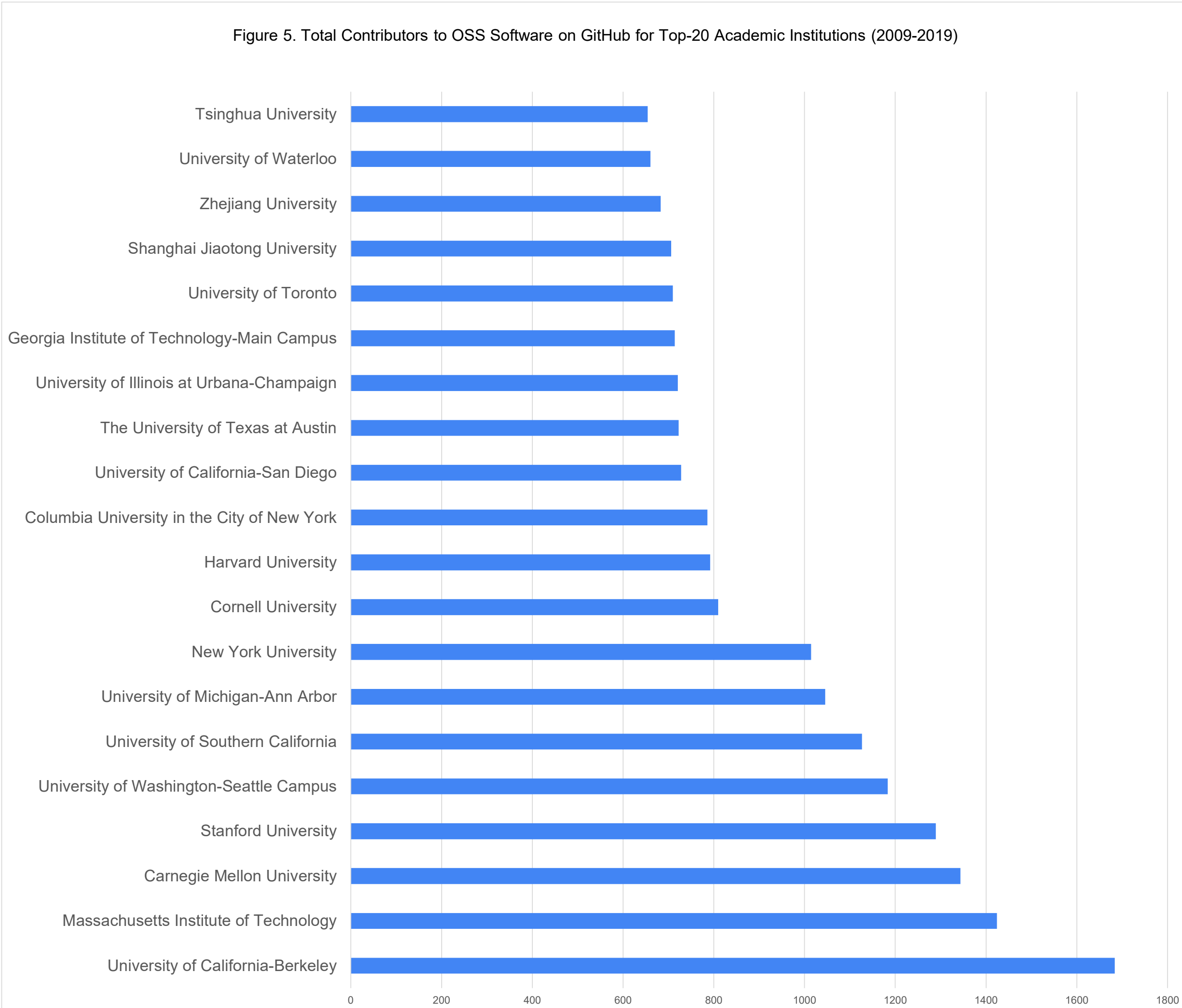


Figure 5. Total Contributors to OSS Software on GitHub for Top-20 Academic Institutions (2009-2019)
Tsinghua University
University of Waterloo
Zhejiang University
Shanghai Jiaotong University
University of Toronto
Georgia Institute of Technology-Main Campus
University of Illinois at Urbana-Champaign
The University of Texas at Austin
University of California-San Diego
Columbia University in the City of New York
Harvard University
Cornell University
New York University
University of Michigan-Ann Arbor
University of Southern California
University of Washington-Seattle Campus
Stanford University
Carnegie Mellon University
Massachusetts Institute of Technology
University of California-Berkeley
0
200
400
600
800
1000
1200
1400
1600
1800