Journal of Data Science 0 (0), 1–20
??? 2024
DOI: 10.6339/24-JDS1148
Data Science in Action

Measuring Public Open-Source Software in the Federal Government: An Analysis of Code.gov

Rahul Shrivastava[1] and Gizem Korkmaz[1,*]
[1]*Statistics and Data Science Center, Westat, Rockville, MD 20850, United States*

Abstract

This paper presents an in-depth analysis of patterns and trends in the open-source software (OSS) contributions by the U.S. federal government agencies. OSS is a unique category of computer software notable for its publicly accessible source code and the rights it provides for modification and distribution for any purpose. Prompted by the Federal Source Code Policy (USCIO, 2016), Code.gov was established as a platform to facilitate the sharing of custom-developed software across various federal government agencies. This study leverages data from Code.gov, which catalogs OSS projects developed and shared by government agencies, and enhances this data with detailed development and contributor information from GitHub. By adopting a cost estimation methodology that is consistent with the U.S. national accounting framework for software investment proposed in Korkmaz et al. (2024), this research provides annual estimates of investment in OSS by government agencies for the 2009–2021 period. The findings indicate a significant investment by the federal government in OSS, with the 2021 investment estimated at around $407 million. This study not only sheds light on the government's role in fostering OSS development but also offers a valuable framework for assessing the scope and value of OSS initiatives within the public sector.

Keywords *Code.gov; cost measurement; Github; open-source software; software investment*

1 Introduction

Open-source software (OSS) is a computer software whose source code can be accessed, used, modified, and shared publicly. OSS is often distributed under licenses that comply with the definition of "open source" provided by the Open Source Initiative (OSI, 1998). The philosophy behind open source goes beyond mere access to the code; it fosters a sense of community and shared progress among developers, who can contribute to projects, improve existing programs, and share knowledge and techniques. The contribution to OSS is made by a variety of entities, including universities, businesses, government research institutions, and nonprofits. Widely-used examples include Apache which was developed with federal and state funds at the National Center for Supercomputing Applications in Illinois, Linux which was first developed at the University of Helsinki, and the R language which was developed at the University of Auckland in New Zealand by professors for use in their teaching laboratory, with extended development at Rice and Stanford Universities (Robbins et al., 2018a,b). Additionally, substantial contributors from the U.S. federal government include the Sandia National Laboratory (SNL), the Lawrence Livermore National Laboratory, and the General Services Administration (Hoffa, 2017). The annual invest-

*Corresponding author. Email: gizemkorkmaz@westat.com.

Received December 9, 2023; Accepted July 5, 2024

ment in software as a share of the U.S. total investment in fixed assets has doubled from 6.2% in 1997 to 12.8% in 2022.[1] In 2022, annual investment in software is estimated at $702 billion ($598 billion from the private sector and $104 billion from the public sector) (Korkmaz et al., 2024).

Recent policies of the U.S. federal government such as the Federal Source Code Policy (FSCP) promote the sharing of software source code developed by or for the federal government (Scott and Rung, 2016). This policy provides a framework for government code to be released and reused through open-source software 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 (Robbins et al., 2021). As of June 2022, more than 8,500 software projects are shared on Code.gov by twenty-one federal departments, agencies, and other entities for reuse by anyone. Other actions and initiatives taken by the U.S. government, such as the Foundations for Evidence-Based Policymaking Act of 2018 (Congress, US, 2018), the Executive Order 13960 on promoting the use of trustworthy artificial intelligence in the federal government (Biden, 2020), the Executive Order 14110 on safe, secure, and trustworthy development and use of artificial intelligence (Biden, 2023), promote the transparency and documentation of methods, including software.

Despite its widespread use and significant role in various sectors, there is a notable lack of reliable metrics to quantify the scope and value of OSS, particularly those projects developed with public funds. Current and comprehensive survey data do not exist for the contributions of OSS (Keller et al., 2018; Robbins et al., 2018b; Korkmaz et al., 2024). There is no readily available accounting of OSS created and funded by government agencies (Robbins et al., 2018b). Public investment in software is published in the national accounts, and the investment of software from the federal government, excluding Department of Defense, is allocated to nondefense account based on the National Income and Product Accounts (NIPA) Handbook (US Bureau of Economic Analysis, 2022). However, the current classification system is limited as it does not distinguish between open-source and proprietary software investment in the federal government (Korkmaz et al., 2024). Additionally, there is a lack of knowledge about the impact and prevalent use of OSS created by the government. As the U.S. federal government agencies share software through well-known platforms, such as GitHub, 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 national laboratories, such as SNL, a wealth of information (contributors, lines of code, organizations) is available in the source code repositories.

Our aim is to measure the scope of OSS developed in the U.S. federal government and estimate its value (measured by investment and use). In this paper, we acquired a list of OSS projects cataloged by agencies from Code.gov along with the projects' descriptions and online locations of the hosting repository, and we gathered additional development information from GitHub. We measure OSS contributions by the government agencies, and explore top organizations and sectors contributing to projects listed on Code.gov. We find that the Department of Energy (DOE), the General Services Administration (GSA), the Department of Health & Human Services (HHS), and NASA are consistently among the top contributing agencies in the federal government based on the number of repositories and development activity.

Our analysis provides insights into how federal agencies are implementing policies following the FCSP and documents the rate at which federal agencies are adopting OSS. Agencies are

[1] See https://alfred.stlouisfed.org/graph/?g=1bFiA for the calculation using data from the U.S. Bureau of Economic Analysis (BEA) (US Bureau of Economic Analysis, 2023a,b,c,d,h,i,j).

required to perform various tasks in order to satisfy the objectives of the FCSP, however, agency compliance varies (see dashboard on Code.gov (n.d.)). Most agencies (except for the Department of Defense (DOD) and the Department of Interior) have updated or are updating their policies for consistency with FCSP based on the compliance dashboard on Code.gov (n.d.). Some agencies have already developed an inventory of all custom-developed software and reported this inventory to Code.gov (e.g., software policies of DOE (DOE CODE, n.d.), GSA (GSA, 2019)). Agencies use a combination of site-maintained repositories, private repositories (such as site-hosted GitLab), and public repositories (such as GitHub and Bitbucket). For example, NASA asks employees to create a GitHub account and to submit a request to be added to NASA's GitHub page (NASA, n.d.). We observe differences in the implementation of the FCSP policy. The majority (70%) of open-source projects listed by all agencies on Code.gov are hosted on GitHub (76% of DOE's 3,760 projects, 27% of NASA's 1,276 projects, and all projects listed by GSA (1,168 projects), HHS (583 projects), Department of Homeland Security (DHS) (351 projects), Veterans Affairs (VA) (148 projects), and DOD (43 projects)). The Department of Transportation (USDOT), DOE and NASA share a large amount of projects on .gov sites (84% of USDOT's 132 projects, 70% of NASA's projects, and 15% of DOE's projects, mainly those from national laboratories, are hosted on .gov sites) – these sites usually require user registration to obtain the software, or the URL's provide compressed files for download. Some agencies seem to be slower in the implementation of the OSS policies, i.e., in posting their inventory with complete information to Code.gov. For example, the Department of Housing and Urban Development (HUD) lists 172 projects, but does not provide any URL's to the code or to the repository of these projects (even though they are compliant based on the dashboard). Similarly, 95% of the Social Security Administration's listed 132 projects and almost 50% of the Department of Labor's 178 projects do not have a repository location. The National Science Foundation provides the URL of main web page (nsf.gov) as the repository for 18 of their listed 32 projects. This lack of information results in an underestimation of the federal government's OSS investment provided in this paper.

To explore the affiliations of the contributors to the federal government agencies' OSS repositories, we use software packages diverstidy (Kramer, 2021a) and tidyorgs (Kramer, 2021b) and assign the GitHub users to countries and sectors based on the self-reported location and organization information along with the email addresses. We find that the OSS projects listed on Code.gov involve contributors affiliated with various countries and sectors in addition to the United States and the government sector. 23% of the contributors are affiliated with the academic sector (top universities include Stanford University, University of California – Berkeley and the University of Michigan – Ann Arbor) and 20.7% with businesses (among the top are Intel, Google, Kitwar and Red Hat).

By adopting a cost estimation methodology consistent with U.S. national accounting framework used for measuring software investment (proposed in Korkmaz et al., 2024), we generate annual estimates of investment in OSS by government agencies for the 2009–2021 period. The annual investment in OSS in 2021 is estimated as approximately $407 million. In 2021, about 76% of investment came from DOE. Based on the impact of the projects (measured by forks and stars, described in Section 4.4), DOD, Treasury and the Federal Election Committee are found as being highly influential even though they are not among the top agencies based on previous production-based metrics. This highlights the importance of considering both demand and supply side when valuing OSS projects.

This paper is organized as follows. Section 2 summarizes relevant literature, including the economic measurement of software and other intangibles. Data and methods for this paper are

described in Section 3 and results are presented in Section 4. Section 5 concludes by providing a discussion of limitations and extensions to this work.

2 Related Work

Many types of innovation often represent intangible assets that are hard to measure, such as knowledge and open-source software (Damanpour, 1991). This "dark innovation" (Martin, 2016; Keralis et al., 2023) takes place in households, universities, and governments, and occurs when the product is used, rather than sold in the market (Gault, 2018), and is referred to as free innovation (Von Hippel, 2016) or household production (Bockstael and McConnell, 1983). Traditional measures of innovation (using surveys) do not fully capture the value of these assets. Not valuing these intangibles misses changes in the economy, and it leads to underestimation of productivity and misallocation of resources. Better accounting of public investment in intangibles would provide a more complete picture of economic growth (Corrado et al., 2015).

Keller et al. (2018) uses OSS innovation as a case study, and describe the challenges and processes to measure these intangibles using a data science framework. Through a process of data discovery, acquisition, statistical data integration, and visualization, the authors show the feasibility of measuring innovation related to OSS through data collected from online software repositories. They provide evidence and insights about how these data could be used to estimate value and impact. Similarly, our approach is to observe and measure intangible inputs to innovation using non-survey data sources (e.g., software repositories).

National economic accounting methods have adjusted to the increasing interest in better measurement of the economic impact of computer software and the increased digitization of knowledge. In 2019, annual investment in software in the U.S. is estimated at $527 billion, with $104 billion (or 15%) from the public sector which includes federal, state and local governments (see published series (US Bureau of Economic Analysis, 2023a,b,c,e,f,g)). The current methodology has limitations for measuring OSS investment as it does not distinguish between open-source and proprietary software. Given the different implications for productivity between these types of software, it is vital to provide separate measures.

In the absence of a price, intangible assets can be valued based on their production cost (Nakamura and Soloveichik, 2015; Nakamura et al., 2017). A production cost based approach to measure OSS as intangible capital created within and outside of the business sector (such as in universities, federal government agencies, and households) using non-survey data was initially prototyped in Robbins et al. (2018a,b) for four open-source programming languages: R, Python, Julia, and JavaScript. The authors estimated that the resource cost for developing R, Python, Julia, and JavaScript exceeds $3 billion dollars, based on 2017 costs.

Robbins et al. (2021), Calderón et al. (2022), and Korkmaz et al. (2024) extended the previous work to 7.75 million GitHub repositories – all projects (with an OSI-approved license) developed on GitHub. The authors generate annual estimates of U.S. investment (nominal and real) in OSS that are consistent with measures of software investment in the national accounts used by the BEA. Their estimates show that the U.S. investment in OSS in 2019 was $37.8 billion, and the portion for the government is estimated at $445 million. Our paper builds on the current national account framework and cost methodology proposed in Korkmaz et al. (2024), however differs in two major ways. Korkmaz et al. (2024) use Code.gov and GitHub data and developer information to identify users affiliated to federal government agencies. They are able to assign a sector to 12% of all users in the data, and 0.1% is assigned to the government

sector – which is the basis for the investment estimates. They also identify and include additional GitHub repositories these users contribute to (even if the projects are not listed on Code.gov). Our approach is to use all projects listed on Code.gov and estimate the production cost based on the lines of code added to each GitHub repository. Moreover, we are interested in measuring the impact of these projects (demand side) in addition to the cost estimates (supply side).

Hoffmann et al. (2024) measures the value of OSS, for both demand and supply side, and find a value ranging from $1.22 billion to $6.22 billion for all widely-used OSS. The authors focus on the code that goes into products that firms create, and the code that consumers directly interact with through firm websites, while our goal is to estimate the value of OSS developed by the federal government agencies. For the demand-side, the authors use a labor market approach and estimate the entire cost (for recreation of all OSS used by companies) to be between $2.59 trillion to $13.18 trillion. In this paper, we do not estimate the demand side value, however, we use GitHub metrics, forks and stars, to measure the impact of OSS projects developed and shared by the federal government agencies.

Researchers have been stressing the importance of quantifying coding impact in the scientific community (Howison et al., 2015; Singh Chawla, 2016). Traditional evaluation of the academic work is built around the citations which are used to measure the impact of the research outputs of universities and government institutions such as patents and publications (Garfield et al., 2002; Science-Metrix, 2018; Rehn et al., 2014). However, few academic papers actually cite software (Piwowar and Priem, 2016). Howison and Bullard (2016) found that two-thirds of 90 randomly selected biology papers mentioned the software used, but less than half of them actually cited the package. Even though developers write papers that describe their software, researchers may not know which paper to cite because software packages often have multiple articles associated with them (Singh Chawla, 2016). This results in a lack of citation information for OSS projects (Keller et al., 2018; Korkmaz et al., 2018) which limits measuring the impact of this innovation.

To quantify coding impact of software built by academics, researchers built Depsy.org (Piwowar and Priem, 2016) in 2015 with NSF-funding, compiling R and Python packages (Impact Story, 2012). Korkmaz et al. (2018, 2020) present an impact-focused approach that uses the number of downloads and citations of OSS as measures of their impact. Using data collected from Depsy.org, the authors generate dependency and contributor networks of R and Python packages and develop statistical models to identify factors that affect the impact of OSS. In this paper, we use network centrality metrics to measure impact of developers and OSS metrics (forks and stars) as measures of project impact.

3 Data and Methods

Many OSS projects are developed and maintained in free repositories. Hosting platforms which are based on Git version control, such as GitHub and GitLab, are used to develop, download, review, and publish code for projects that are stored and managed in central locations referred to as repositories. These platforms host public and private repositories while providing a suite of features such as access and permissions by teams, issue tracking, wikis, web-hosting, and continuous integration. Version control systems, such as Git, serve for tracking changes and coordinating work on files among multiple developers. Information embedded in the repositories and websites, including the code, contributors, and development activity, is publicly available through the use of Application Programming Interface (API) and web-scraping, and creates a

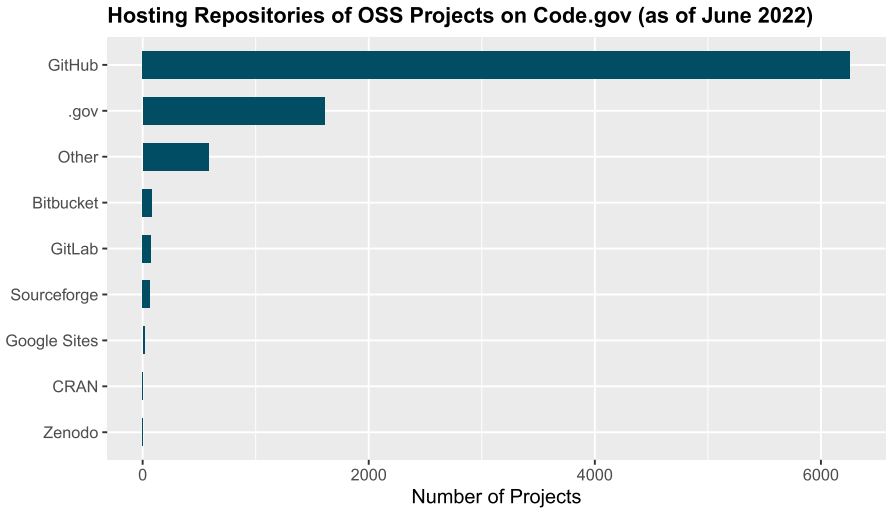

Figure 1: Distribution of repository hosting platforms used by federal government agencies.

very rich source of data to study the scope and impact of these projects (Keller et al., 2018). This paper shows how these data can be used to measure OSS innovation by the U.S. federal government agencies. In this section, we describe the data and methods used to estimate the total costs involved in the production of software and the impact of the projects.

3.1 Data

We acquired data from Code.gov in JSON format that included 8,697 software projects across 21 different federal agencies as of June 2022. The data set includes information on each project such as agency name, URL of the hosting repository, and description. The JSON files were downloaded from the Code.gov platform and Python libraries, pandas (pandas, 2020) and NumPy (Harris et al., 2020), were used to extract project information and hosting repository links.

Repositories shared on Code.gov are hosted on a variety of platforms including GitHub, Code.gov, Bitbucket, GitLab, Sourceforge, Google Sites and others. Figure 1 illustrates the use of different hosting platforms where the code is developed and maintained – GitHub is the most popular – 69.5% of all repositories listed on Code.gov are developed and hosted on GitHub. This includes 76% of DOE's projects, 27% of NASA's, and all projects listed by GSA, HHS, DHS, DOD, and VA. USDOT, DOE and NASA share a large amount of projects on .gov sites (84% of USDOT's, 70% of NASA's projects, and 15% of DOE's projects, mainly those from national laboratories, are hosted on .gov sites) – these sites usually require user registration to obtain the software, or the URL's provide compressed files for download. Bitbucket, GitLab, Google Sites, CRAN and Zenodo are provided as repository location only by DOE's projects, and they make up of less than 1% of DOE's projects. Sourceforge is listed by DOE for 27 projects, and by NASA for 40 of their projects.

Our second source of data is GitHub – which is the largest source code hosting platform in the world with over 100 million developers and 284 million public repositories (GitHub, 2023). To collect the data from GitHub we utilized the GitHub GraphQL API[2] to obtain information on

[2]GitHub's GraphQL API enabled faster and more efficient data collection over the REST API. GraphQL

repositories, commits and individual users. Data collected from GitHub was parsed and stored in a MySQL Database on an Amazon AWS RDS Database instance.

We collected publicly available metadata on the 6,047 repositories corresponding to Code.gov projects hosted on GitHub along with information on 35,503 user accounts associated with 17,577 unique contributors (usernames/logins). The information gathered from GitHub includes commits (approved code changes), lines of code added and deleted, license detail, users' login information, organization, location, and email.

Finally, we used software packages Kramer (2021a) (a tidy package for detection and standardization of geographic, population, and diversity-related terminology in unstructured text data) and the Kramer (2021b) (a tidy package that detects and standardizes organizations in unstructured text data) to obtain further information on developers such as their organizations, countries, and sectors (business, academic, government, nonprofit).

3.2 Cost Estimation Methodology

We adopt the OSS cost estimation methodology proposed in Korkmaz et al. (2024). The authors use a close adaptation of the basic Constructive Cost Model (COCOMO) (Boehm, 1984), which was originally developed for estimating the cost and effort of developing proprietary software projects, and was designed to be flexible and adaptable to various software development projects. The intuition behind the model is that the required effort and development time for software can be approximated by observable factors such as the size of the project, which can be measured in lines of code. The estimates of nominal government labor investment per year are computed from data on the number of lines of source code changed each year, the labor cost of a programmer's time, and estimates of how many lines of code programmers edit per hour. The model provides a calibration factor — the effort required in person-months to produce a set number of lines of code — and a range of parameters (effort multipliers) that can be used to adjust for the complexity of the project. The adaptation of COCOMO in Korkmaz et al. (2024) takes the following functional form and parameters:

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

$$\text{Nominal development time} = 2.5(\text{Effort})^{0.38}$$

$$\text{Annual development cost} = (\text{Monthly Resource Cost})\ (\text{Nominal development time})$$

KLOC stands for kilo (thousand) lines of code added per year and the current model treats each line equally (i.e., the model does not adjust for programming languages or complexity of the code). The effort in person-months is estimated by multiplying lines of code of the project by the calibration factor and the effort multiplier (Boehm, 1984). Effort multipliers account for complexity, reliability, and scale for these models. We use the parameters from the basic COCOMO corresponding to the organic software class, which consists of software dealing with a well-known programming language and a small, but experienced team of contributors. The resulting nominal development time in months is then multiplied by the estimated person-month production cost based on monthly average wages for programming occupations adjusted with a factor to account for the non-wages components (e.g., additional labor costs, capital services, intermediate inputs) of the total cost of production.[3] Lastly, we apply a price deflator to adjust

allows for specifying exact data endpoints and enables collection of information using fewer API requests.

[3]The wage series and blow-up factors are consistent with those of the own-account software methodology (US Bureau of Economic Analysis, 2022). The average wages for the three occupations considered for own account soft-

Table 1: Network measures and definitions.

Network Measure	Definition
Degree (weighted)	Total number of links (takes into account multiple links between nodes)
Diameter	Shortest distance between the two most distant nodes in the network
Components	Portions of the network that are disconnected from each other
Clustering coefficient	Degree to which node's neighbors are connected
Degree centrality	Normalized by dividing by the number of nodes
Betweenness centrality	Number of shortest paths that pass through a node

for inflation and quality changes over time which yields the real investment series (Korkmaz et al., 2024).

3.3 Generating Contributor Networks

OSS interactions are complex and involve multiple types of actors and various interaction types. The analysis of the structural features of the networks can help us characterize the patterns and dynamics of collaborations between individual developers, within and across institutions, sectors, and countries. In this paper, we define and generate contributor networks for each agency and analyze structural features of these networks. A Contributor Network is generated by extracting undirected edges between individual contributors (GitHub users) each time they contribute to the same project (repository) with edge weights measuring frequency of collaborations.

Nodes: A node represents each individual contributor or GitHub user contributing to repositories on GitHub associated with Code.gov software projects.

Edges: An edge between two developers is formed if they contribute to the same GitHub repository associated with a Code.gov project.

Weights: The weight of an edges represents the number of GitHub repositories two developers contribute to. For example, if two individual developers (GitHub users) are contributors to 3 repositories then the weight of the edge between them will be 3 in the network.

We created edge lists corresponding to each agency and characterize the networks using (i) measures for properties of nodes and edges, such as centrality measures (listed in Table 1 below), (ii) local measures that describe the neighborhood of a node or the occurrence of subgraphs (e.g., clustering), and (iii) global measures analyzing the interconnectivity structure such as number of components, and diameter (Börner et al., 2007).

4 Results

In this section, we summarize our findings based on the linked Code.gov and GitHub data. In Section 4.1, we present contributions to OSS by the government agencies, and explore top countries, sectors, and organizations contributing to projects listed on Code.gov. In Section 4.2,

ware in the business sector were obtained from the Occupational Employment and Wage Statistics (OEWS) (US Bureau of Labor Statistics, 2021). A blow-up factor of 2.02 is used based on the multi-year total expenses to gross payroll ratio based on the Services Annual Survey (SAS) data for the representative industry Computer Systems Design and Related Services (NAICS: 5415) (Calderón et al., 2022; Korkmaz et al., 2024).

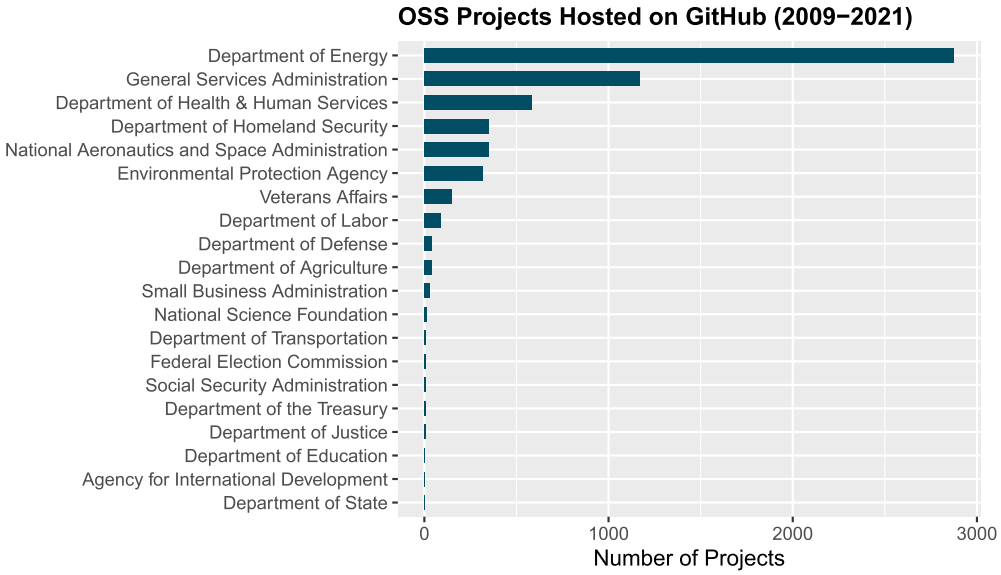

Figure 2: GitHub repositories created by federal government agencies between 2009 and 2021.

we provide estimates of investment in OSS measured by the cost of development. We demonstrate the relative sizes of investment in OSS by agencies between 2009 and 2021. In Section 4.3, we present the structural features of contributor networks generated for the federal government agencies. Finally, in Section 4.4, we switch to the demand/impact side, and demonstrate the impact of projects developed by the agencies using software metrics – forks and stars.

4.1 Measuring the Scope of OSS in the Federal Government

We analyze the top contributing agencies (measured by number of repositories) and the contributor affiliations (countries, sectors, and organizations). As mentioned in Section 3, there are 6,047 repositories listed on Code.gov that are hosted on GitHub. Figure 2 shows the number of GitHub repositories by government agencies. DOE has the highest number of GitHub repositories (2,876 projects) followed by GSA with 1,168 projects and HHS with 583 projects. Korkmaz et al. (2024) present a summary of OSS contributions on GitHub by all users with institutional emails of the federal government.[4] The authors calculate the total number of GitHub repositories contributed by these users between 2009–2019 as 15,716. Consistent with our findings, the top contributors based on the number of repositories are found as DOE, NASA, and HHS.

Second, we focus on the contributors of these repositories hosted on GitHub, and analyze the characteristics using their self-reported affiliations and locations. Table 2 illustrates a breakdown of users' assignment to country and sectors. Note that not all contributors provide complete information in their profiles, hence we were not able to map those contributors to countries or sectors. Using the software diverstidy (Kramer, 2021a) and tidyorgs (Kramer, 2021b), 51% of the users were assigned to at least one country, and 31% were assigned to a sector based on the

[4]The authors obtain a list of OSS projects from Code.gov and identify their contributors on GitHub that are affiliated with the U.S. federal government. They also identify and include additional GitHub repositories these users contribute to (even if the projects are not listed on Code.gov).

Table 2: Country and sector assignment to contributors (2009–2021).

	Total	% All	% Valid
Total Contributors	17,577	100	–
Any Country	9,226	51.4	100
United States	4,916	27.9	53.3
Germany	656	3.7	7.1
United Kingdom	588	3.3	6.4
France	313	1.8	3.4
Canada	279	1.6	3.0
Multiple countries	648	3.7	7.0
In Any Sector	5,458	31.0	100
Government	2,720	15.5	49.8
Academic	1,254	7.1	23.0
Business	1,132	6.4	20.7
Nonprofit	144	0.1	2.6
Multiple sectors	208	1.2	3.8

self-reported location and organization information along with the email addresses. As expected, the majority (53.3%) of the users (of those that were assigned to a country) is affiliated with the United States. Other developers contributing to federal government agencies' repositories are associated with countries such as Germany, United Kingdom, France, and Canada.

Moreover, OSS projects listed on Code.gov involve contributors affiliated with various sectors, such as business, academia, and nonprofits. After the government sector (49.8%), contributors affiliated with academia (with 1,254 contributors (23%)) are the second most prominent group to contribute to Code.gov projects on GitHub. 1,132 contributors (20.7%) were affiliated with the business sector, and 144 contributors (2.6%) with the nonprofits.

Figure 3 illustrates top organizations these contributors are affiliated with. The highest number of individual contributors were from the LLNL and SNL. Top organizations also include businesses – Intel Corporation and Google. Moreover, Figure 4 illustrates top organizations in the business sector that contribute to Code.gov projects on GitHub. Intel Corporation, Google, Kitware, Red Hat, and Nvidia are among the top organizations with the highest number of contributors to Code.gov projects.

Finally, we take a deeper dive into the academic sector, and analyze the top universities affiliated with the contributors. Figure 5 illustrates top 20 universities (U.S. and foreign). We observe that Stanford Univ. has the highest number of contributors (227 users) followed by the Univ. of California – Berkeley (153 users) and the Univ. of Michigan – Ann Arbor (97 users). This closely aligns with Korkmaz et al. (2024) that analyzes the affiliations of around 105 thousand GitHub users (out of 3.2 million) with self-reported university affiliations.

4.2 OSS Investment by the U.S. Federal Government Agencies

In this section, we present results based on methods described in Section 3. Table 3 presents the estimates of the real investment in OSS for top 10 agencies (and the total for all agencies) in 2021 dollars using the own account software price index (Korkmaz et al., 2024). As a reference,

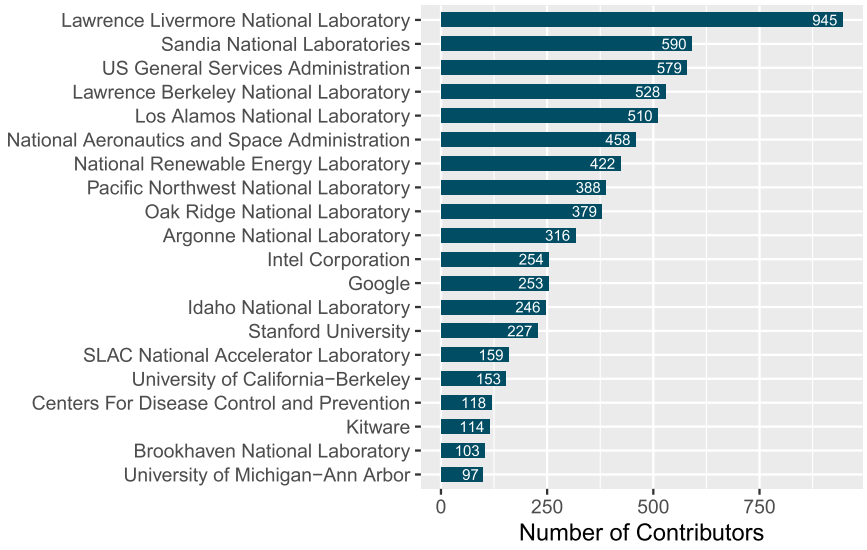

Figure 3: Top organizations associated with GitHub users contributing to projects on Code.gov.

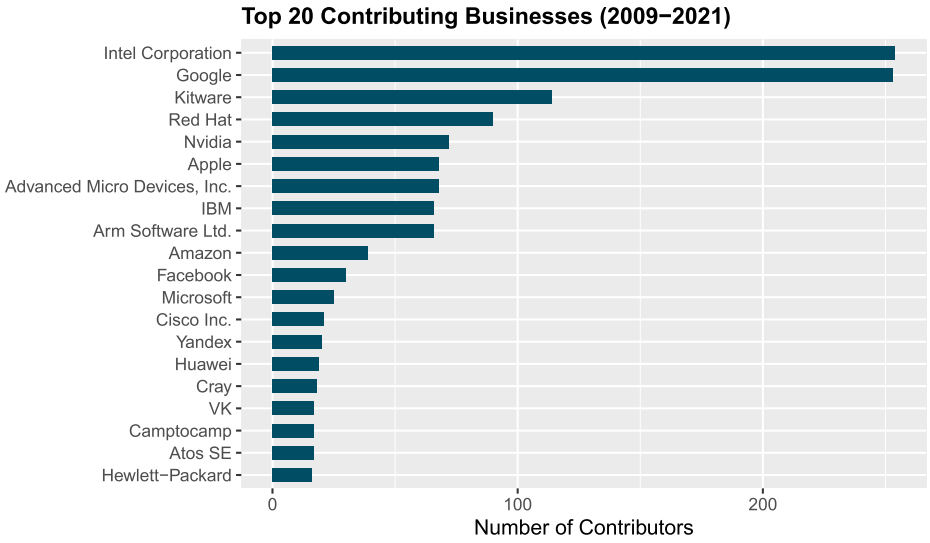

Figure 4: Top organizations in the business sector contributing to projects on Code.gov.

the last column shows KLOC added to agencies' repositories in 2021. The annual investment in OSS in 2021 is estimated at approximately $407 million. On average, investment by each agency is increasing over time. In 2021, about 76% of investment came from DOE corresponding to $308 million – which is significantly higher than the investments by other agencies.

4.3 Analysis of the Contributor Networks

4.3.1 Agency Networks

To capture the differences in collaboration dynamics within and across federal government agencies, we generate the contributor networks where an undirected edge between i and j indicates

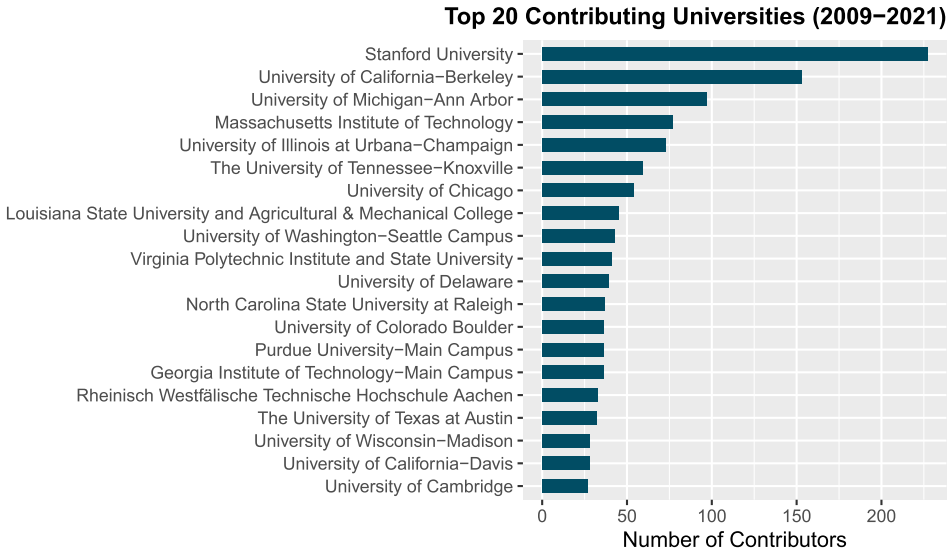

Figure 5: Top universities associated to GitHub users contributing to projects on Code.gov.

Table 3: Investment (in million dollars) by U.S. federal government agencies.

	2009	**2012**	**2015**	**2018**	**2021**	**KLOC**
Department of Energy (DOE)	20	56.6	128.6	256.9	308	2,138,677
Department of Homeland Security (DHS)	–	–	0.5	5.7	30	21,888
General Services Administration (GSA)	1.8	9	20.4	34.6	22	153,490
National Aeronautics and Space Administration (NASA)	1.6	4.7	21.5	17.8	18	169,199
Department of Health and Human Services (HHS)	0.2	2.2	2.5	13.5	15	43,263
Department of Defense (DOD)	–	0.4	2.7	4.7	5	30,824
Department of Agriculture (USDA)	–	–	2.2	7.1	2	28,150
Small Business Administration (SBA)	–	–	–	0.9	1	1,829
Department of Justice (DOJ)	–	–	–	0.8	1	2,473
Veterans Affairs (VA)	–	–	0.5	1.2	1	8,025
Total (21 agencies)	**23.6**	**73.7**	**184.9**	**351.5**	**407**	2,641,616

Note: The investment is adjusted for inflation across time, calculated by dividing current investment by a price deflator. The price deflator used is the price index for private fixed investment in intellectual property products: Software: Own account [Y005RG3A086NBEA], Vintage: 2023-12-01.

that users i and j contribute to the same repository (see definitions in Section 3.3). The full network consists of 17,185 nodes/users and 2.57 million edges/collaborations, and is composed of 233 connected components. The giant component has 16,166 nodes, corresponding to 94% of the full network. Next, we subset the contributor network by each agency and analyze the properties of each agency's network. Table 4 summarizes the structural properties of these networks (for top 5 based on the size of the network). The average degree of the network is calculated by dividing the total degree (unweighted) by the total number of nodes.

We observe that these networks are structurally quite different although for all of them the average degree is higher than other collaboration type networks. We obtain the largest network for DOE with over 10,000 nodes (collaborating developers) and 1,635,040 edges. The contributor network for GSA is much smaller with approximately 4,000 nodes and the number of

Table 4: Structural features of the contributor networks.

	Nodes	Edges	Avg. degree	Diameter	Avg. clustering coefficient	Connected components
Dept. of Energy	10,332	1,635,040	316.50	10	0.92	154
General Services Administration	3,967	806,880	406.79	7	0.95	29
National Aeronautics and Space Admin.	1,202	41,975	69.84	8	0.92	36
Dept. of Homeland Security	500	11,669	46.67	5	0.95	6
Dept. of Health and Human Services	472	6,707	28.41	7	0.95	15

connections are half as much as the DOE network. This implies that there is a significantly high level of collaboration on projects associated with GSA. The high average degree of the DOE and GSA networks appear to be a result of a few projects with a large number of contributors. For example, DOE's scipy/scipy repository has 1,301 contributors (each contributor has a degree of at least 1,300 in this network), followed by SOLLVE/llvm with 796 contributors. GSA's largest repository is GSA/terraform with 1,068 contributors followed by GSA/packer which has 570 contributors. In addition, DOE's contributor network is highly disconnected (with over a hundred components) compared to the other agencies networks. It implies that there are fewer collaborations across different DOE projects compared to GSA.

Figures 6 and 7 present the distributions of users' centrality metrics – degree and betweenness – in these networks (sorted by median) to further illustrate differences across agencies. Centrality values are normalized by dividing by the maximum possible degree (i.e., $(n-1)$ where n is the number of nodes). We observe that the contributor networks of USDOT and DOD have highly influential individuals based on these centrality metrics. These contributors have a high level of collaboration and, they act as "bridges" connecting different parts of the networks (betweenness centrality). Note that DOL is among the top agencies based on the median of the nodes' betweenness centrality (see Figure 7) despite its lower rank based on degree centrality (Figure 6). Use of different centrality measures is important to measure the influence of contributors and projects.

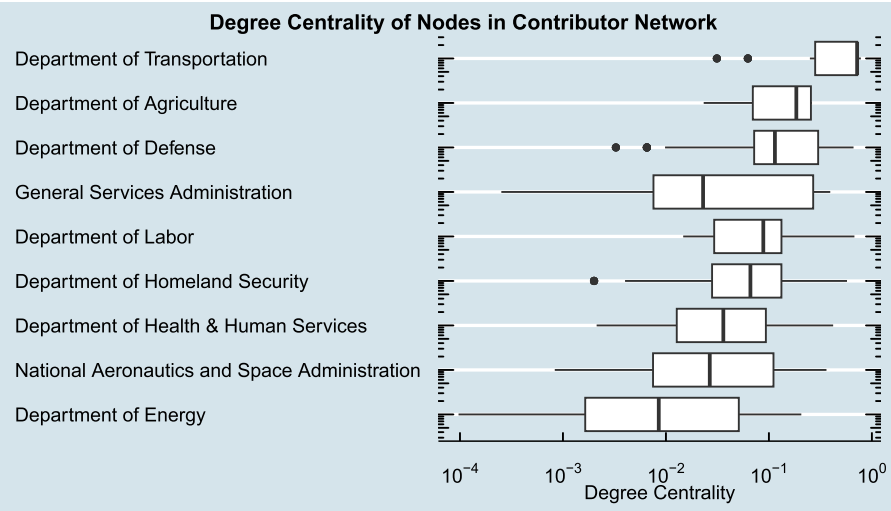

Figure 6: Distribution of users' degree centrality (normalized) in contributor networks.

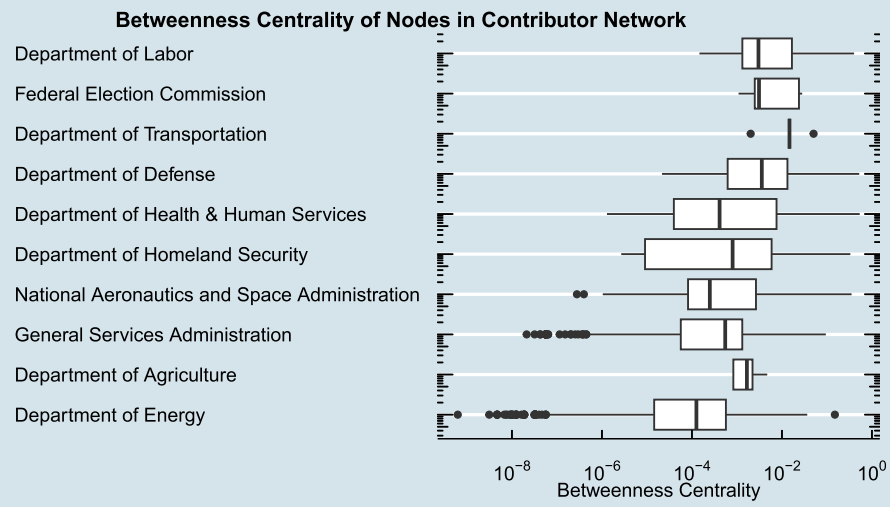

Figure 7: Distribution of users' betweenness centrality (normalized) in contributor networks.

4.3.2 Inter-agency Network

To capture the interaction between agencies, we generate the inter-agency project network where an undirected edge between agency *a* and *b* indicates that agency *a* and *b* have repositories with shared contributors (and the weight measures total number of shared contributors between agencies). In our dataset, there are 364 users that contribute to more than one agency's repository. We obtain a network of 16 agencies (NSF and SBA do not have any connections to (shared contributors with) other agencies). The network is highly connected, with an average degree of 14 and a diameter of 2. Figure 8 illustrates the inter-agency network for the U.S. federal government agencies (plotted using Gephi (Bastian et al., 2009)). The size of the node indicates the weighted degree centrality (e.g., total number of shared contributors an agency has with the other agencies), and the color represents the betweenness centrality of the agencies (the darker orange indicates a higher centrality measure). We observe that GSA, DOE, DHS, HHS, and NASA have the highest number of shared contributors, respectively. These agencies also have high betweenness centrality values, as well as the Dept. of the Treasury, DOL, and VA.

4.4 Measuring the Impact of OSS Projects by the Federal Government

As illustrated in the sections above, based on the number of repositories, development activity and resource cost, and the size of collaboration networks, DOE, GSA, HHS, DHS and NASA are consistently among the top contributing agencies in the federal government. This finding is based on the development (supply) side of the OSS activity.

We are also interested in measuring the use and impact of these projects (demand side), which is important to consider for valuing these projects. The GitHub data include OSS metrics such as forks and stars that can be used to measure impact of these projects. A fork creates a completely independent copy of Git repository which allows the users (beyond the original developer) to make modifications and contributions without altering the original project. Similar to a paper being cited, the higher the number of times a repository is forked, the more its impact will be. The second metric used as a measure of impact is the number of stars which are given

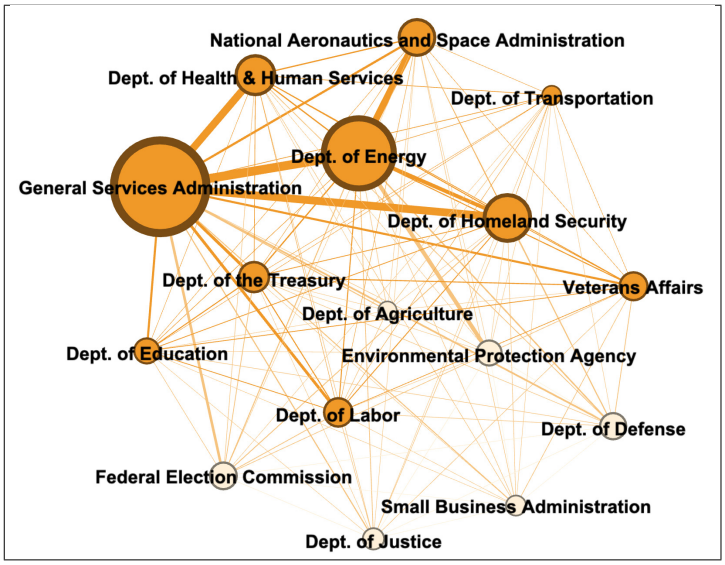

Figure 8: Inter-agency network.

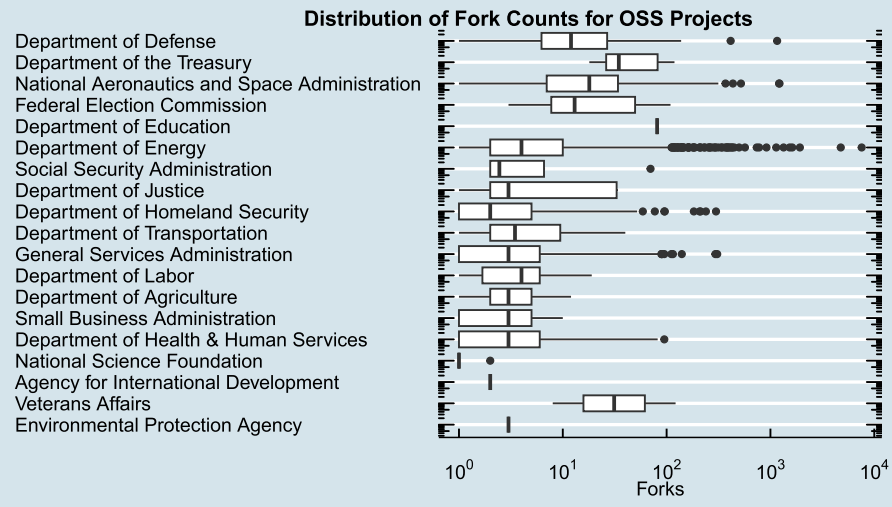

Figure 9: Distribution of fork counts of repositories of the government agencies.

by users when they like a repository and want to show appreciation or to bookmark them so they can follow the developments.

Figures 9 and 10 show the distribution of the counts of forks and stars, respectively, for each project associated with the government agencies. Based on these impact metrics, we observe that DOE projects have a high variation in impact metrics (a number of high impact projects along with medium-impact projects). NASA is still among the top agencies. We observe that DOD, Treasury and the Federal Election Committee have high-impact projects even though they are not among the top agencies based on previous metrics. This highlights the importance of considering both demand and supply side when valuing the OSS projects.

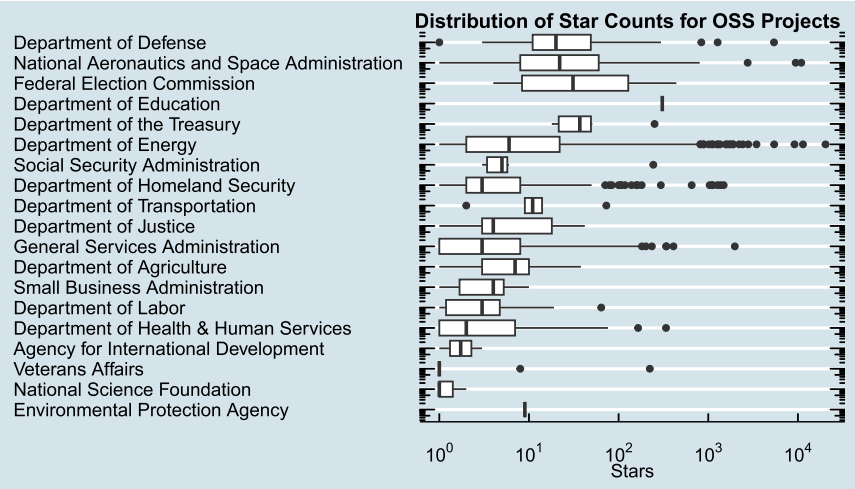

Figure 10: Distribution of star counts of repositories of the government agencies.

5 Conclusion

The integration of OSS in the federal government represents a critical evolution in public sector technology strategy as it accelerates innovation, improves software security and quality, and encourages the democratization of technology, making it accessible and customizable for users around the globe. This study examines the U.S. federal government's involvement in OSS through Code.gov and GitHub, aiming to understand the scope and impact of these initiatives. By analyzing data from these platforms, we seek to identify trends and patterns in government OSS contributions, offering insights into how federal agencies contribute to open-source solutions. The paper also explores the challenges and opportunities presented by this paradigm shift, setting the stage for a deeper understanding of the government's role in the OSS ecosystem.

This paper explores OSS in the U.S. federal government and has highlighted agencies with significant contributions and investments. We provide an economic measure of the cost of production of OSS created by the government and impact measures for its usage and prevalence. The paper highlights agencies with significant contributions and investments, hence documents the rate of adoption of OSS by different federal agencies with the FSCP policy and potential for cost savings with switching to open-source languages.

The paper encounters limitations due to incomplete or inaccurate data: not all government staff contributions appear in this data, because (a) not all government agencies share repositories on Code.gov, and (b) not all development occurs on GitHub. Additionally, contributions of users that are not affiliated with the government agencies are also included in our calculations. Finally, assignment to countries and sectors using self-reported profile information also has its limitations. The provided information may be outdated and may not necessarily reflect their current locations and affiliations.

One of the limitations of the cost estimation methodology (and of existing work, e.g., Korkmaz et al., 2024) is the use of lines of code as a measure of effort. First of all, languages have different standards for the content of the code (some may require more lines of code for the same tasks). Second, lines of code do not account for the quality of the code (e.g., complex functions vs. standard commands). Finally, some repositories include files other than code (such as data files), and this results in overestimation of the effort.

Adjustment of the model based on the 'type' of source code (i.e., weighting the source lines of code by programming language, file type, code complexity) is an important area of future research which could significantly affect the cost estimates. This effort would involve classifying each file in the repository (e.g., .R, .csv, .ipynb, dat, readme) to identify data and code files, as well as developing methods to identify complex functions and simple commands. Moreover, empirical studies are needed to develop language-based multipliers that could be used to convert the line of code to time/effort for each language.

Future work should focus on refining the investment methodology to more accurately identify and weight contributions by federal government employees. This could involve developing methods to verify the sector of employment for GitHub users and adjusting investment estimates accordingly.

Supplementary Material

The data and code needed to reproduce the results in this paper can be found at the Journal of Data Science website. The links to the raw data and code can also be found on our project's website at https://oss.quarto.pub/website/analyses.html.

Acknowledgement

The authors would like to thank anonymous referees for their valuable feedback, and would also like to express their thanks to Ekaterina Levitskaya (Coleridge Initiative) for the initial analysis; J. Bayoán Santiago Calderón (U.S. Bureau of Economic Analysis), Carol Robbins (National Center for Science and Engineering Statistics, National Science Foundation), and Ledia Guci (U.S. Bureau of Economic Analysis) for the cost estimation methodology; and Brandon Kramer (Edge&Node) for their support on utilizing diverstidy and tidyorgs software packages. We are also grateful to GitHub for providing programmatic access to the data through their Application Programming Interface (API).

The earlier versions of the work were presented at the Government Advances in Statistical Programming (GASP) 2023 Conference, and the 35th International Association for Research on Income and Wealth (IARIW) 2018 Conference.

Funding

This work was supported by the National Science Foundation (NSF) under Grant Numbers 2306160 and 2224441. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

References

Bastian M, Heymann S, Jacomy M Gephi: An open source software for exploring and manipulating networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Biden JR (2020). Executive order on promoting the use of trustworthy artificial intelligence in the federal government. https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.

Biden JR (2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Bockstael NE, McConnell KE (1983). Welfare measurement in the household production framework. *American Economic Review*, 73(4): 806–814.

Boehm BW (1984). Software engineering economics. *IEEE Transactions on Software Engineering*, SE-10(1): 4–21. https://doi.org/10.1109/TSE.1984.5010193

Börner K, Sanyal S, Vespignani A (2007). Network science. *Annual Review of Information Science and Technology*, 41(1): 537–607. https://doi.org/10.1002/aris.2007.1440410119

Calderón JBS, Robbins C, Guci L, Korkmaz G, Kramer BL (2022). Measuring the cost of open source software innovation on GitHub. *Technical report*, U.S. Bureau of Economic Analysis.

Code.gov (n.d.). Agency compliance dashboard. https://code.gov/agency-compliance/compliance/dashboard.

Congress, US (2018). Foundations for evidence-based policymaking act of 2018. *Public Law*, 115: 435. https://www.congress.gov/bill/115th-congress/house-bill/4174.

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.

Damanpour F (1991). Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal*, 34(3): 555–590. https://doi.org/10.2307/256406

DOE CODE (n.d.). Software Policy of DOE. https://www.osti.gov/doecode/policy.

Garfield E, Pudovkin A, Istomin V (2002). Algorithmic citation-linked historiography—mapping the literature of science. *Proceedings of the American Society for Information Science and Technology*, 39(1): 14–24. https://doi.org/10.1002/meet.1450390102

Gault F (2018). Defining and measuring innovation in all sectors of the economy. *Research Policy*, 47(3): 617–622. https://doi.org/10.1016/j.respol.2018.01.007

GitHub (2023). The State of the Octoverse. https://octoverse.github.com.

GSA (2019). GSA Open Software Policy. https://open.gsa.gov/oss-policy/.

Harris CR, Millman KJ, van der Walt SJ Gommers R Virtanen P Cournapeau D, et al. (2020). Array programming with NumPy. *Nature*, 585(7825): 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hoffa F (2017). The top contributors to GitHub (2017). https://hoffa.medium.com/the-top-contributors-to-github-2017-be98ab854e87.

Hoffmann M, Nagle F, Zhou Y (2024). The value of open source software. *Harvard Business School Strategy Unit Working Paper* (24-038).

Howison J, Bullard J (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *The Journal of the Association for Information Science and Technology*, 67(9): 2137–2155. https://doi.org/10.1002/asi.23538

Howison J, Deelman E, McLennan MJ, Ferreira da Silva R, Herbsleb JD (2015). Understanding the scientific software ecosystem and its impact: Current and future measures. *Research Evaluation*, 24(4): 454–470. https://doi.org/10.1093/reseval/rvv014

Impact Story (2012). https://impactstory.org.

Keller SA, Korkmaz G, Robbins CA, Shipp SS (2018). Opportunities to observe and measure intangible inputs to innovation: Definitions, operationalization, and examples. *Proceed-*

ings of the National Academy of Sciences, 115(50): 12638–12645. https://doi.org/10.1073/pnas.1800467115

Keralis JM, Albertorio-Díaz J, Hoppe T (2023). Dark citations to federal resources and their contribution to the public health literature. *Frontiers in Research Metrics and Analytics*, 8: 1235208. https://doi.org/10.3389/frma.2023.1235208

Korkmaz G, Kelling C, Robbins CA, Keller SA (2018). Modeling the impact of R packages using dependency and contributor networks. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 511–514.

Korkmaz G, Kelling C, Robbins CA, Keller SA (2020). Modeling the impact of Python and R packages using dependency and contributor networks. *Social Network Analysis and Mining*, 10: 1–12. https://doi.org/10.1007/s13278-019-0612-8

Korkmaz G, Santiago Calderón JB, Kramer BL, Guci L, Robbins CA (2024). From GitHub to GDP: A framework for measuring open source software innovation. *Research Policy*, 53(3): 104954. https://doi.org/10.1016/j.respol.2024.104954

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

Kramer BL (2021b). tidyorgs: A tidy package that standardizes text data for organizational analysis. https://github.com/brandonleekramer/tidyorgs.

Martin BR (2016). Twenty challenges for innovation studies. *Science and Public Policy*, 43(3): 432–450. https://doi.org/10.1093/scipol/scv077

Nakamura LI, Samuels J, Soloveichik RH (2017). Measuring the 'free' digital economy within the GDP and productivity accounts. https://www.bea.gov/research/papers/2017/measuring-free-digital-economy-within-gdp-and-productivity-accounts.

Nakamura LI, Soloveichik RH (2015). Valuing 'free' media across countries in GDP. *FRB of Philadelphia Working Paper*.

NASA (n.d.). NASA Open Software Policy. https://code.nasa.gov/#/guide.

OSI (1998). The open source definition. https://opensource.org/osd.

pandas (2020). pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134.

Piwowar H, Priem J (2016). Depsy: Valuing the software that powers science. https://github.com/Impactstory/depsy-research/blob/master/introducing_depsy.md.

Rehn C, Gornitzki C, Larsson A, Wadskog D (2014). Bibliometric handbook for Karolinska Institutet. *Huddinge: Karolinska Institutet*. https://kib.ki.se/sites/default/files/bibliometric_handbook_2014.pdf.

Robbins C, Korkmaz G, Calderon JBS, Kelling C, Shipp S, Keller S (2018a). Open source software as intangible capital: Measuring the cost and impact of free digital tools. In: *International Monetary Fund (IMF) 6th Statistical Forum: Measuring Economic Welfare in the Digital Age: What and How? International Monetary Fund (IMF)*. https://www.imf.org/en/News/Seminars/Conferences/2018/04/06/6th-statistics-forum.

Robbins C, Korkmaz G, Calderon JBS, Kelling C, Shipp S, Keller S (2018b). The scope and impact of open source software: A framework for analysis and preliminary cost estimates. In: *International Association for Research on Income and Wealth (IARIW) 35th General Conference: The Digital Economy-Conceptual and Measurement Issues*. The International Association for Research in Income and Wealth (IARIW). http://old.iariw.org/copenhagen/robbins.pdf.

Robbins CA, Korkmaz G, Guci L, Santiago Calderón JB Kramer B (2021). A first look at open-

source software investment in the United States and in other countries, 2009–2019. In: *International Association for Research on Income and Wealth (IARIW) ESCoE Conference*. IARIW. https://iariw.org/wp-content/uploads/2021/11/robbins-paper.pdf.

Science-Metrix (2018). Bibliometrics and Patent Indicators for the Science and Engineering Indicators 2018. Technical Documentation. http://www.science-metrix.com/en/methodology-report.

Scott T, Rung AE (2016). Federal Source Code Policy: Achieving efficiency, transparency, and innovation through reusable and open source software. *Office of Mgmt. & Budget, Exec. Office of the President Memorandum*. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2016/m_16_21.pdf.

Singh Chawla D (2016). The unsung heroes of scientific software. *Nature News*, 529(7584): 115. https://doi.org/10.1038/529115a

US Bureau of Economic Analysis (2022). NIPA Handbook: Concepts and Methods of the U.S. National Income and Product Accounts.

US Bureau of Economic Analysis (2023a). Government Gross Investment: Federal: National Defense: Gross Investment: Intellectual Property Products: Software. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. Y053RC1A027NBEA.

US Bureau of Economic Analysis (2023b). Government Gross Investment: Federal: Nondefense: Gross Investment: Intellectual Property Products: Software. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. Y068RC1A027NBEA.

US Bureau of Economic Analysis (2023c). Government Gross Investment: State and Local: Gross Investment: Intellectual Property Products: Software. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. Y072RC1A027NBEA.

US Bureau of Economic Analysis (2023d). Gross Government Investment. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. A782RC1A027NBEA.

US Bureau of Economic Analysis (2023e). Private Fixed Investment in Intellectual Property Products: Software: Custom. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. Y004RC1A027NBEA.

US Bureau of Economic Analysis (2023f). Private Fixed Investment in Intellectual Property Products: Software: Own account. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. Y005RC1A027NBEA.

US Bureau of Economic Analysis (2023g). Private Fixed Investment in Intellectual Property Products: Software: Prepackaged. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. Y003RC1A027NBEA.

US Bureau of Economic Analysis (2023h). Private Fixed Investment: Nonresidential: Intellectual Property Products: Software. Retrieved from ALFRED, Federal Reserve Bank of St. Louis. B985RC1A027NBEA.

US Bureau of Economic Analysis (2023i). Private Nonresidential Fixed Investment [PNFIA]. Retrieved from ALFRED, Federal Reserve Bank of St. Louis.

US Bureau of Economic Analysis (2023j). Private Residential Fixed Investment [PRFI]. Retrieved from ALFRED, Federal Reserve Bank of St. Louis.

US Bureau of Labor Statistics (2021). Occupational Employment Statistics: National industry-specific and by ownership.

USCIO (2016). Federal Source Code Policy. https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/memoranda/2016/m_16_21.pdf.

Von Hippel E (2016). *Free Innovation*. MIT Press.