The Scope and Impact of Open Source Software: A framework for analysis with a focus on R and its impacts.

Carol A. Robbins (NCSES, NSF)
Gizem Korkmas (VTSDAL)
Claire Kelling, (Pennsylvania State University)

Abstract

Open source software is everywhere, both as specialized applications nurtured by devoted user communities, and as digital infrastructure underlying platforms used by millions daily. This type of software is developed, maintained, and extended both within the private sector and outside of it, through the contribution of people from universities, government research institutions, nonprofits, and individuals. This paper proposes and implements a methodology to document the scope and impact of open source software created by US Businesses, Government entities, nonprofits, academic institutions and households, extending existing measures of university research output. A prototype estimate of the cost of developing R packages is presented.

Key words: Open Source Software, Intangibles, Innovation, Network Analysis

Introduction and Contribution

Beginning in the early 1980s, open source projects have provided users with zero cost and freely modifiable software tools. These range from the LaTeX typesetting program introduced initially in 1983, and including Linux, Apache, Mozilla Firefox web browser, and the Wordpress content management system. As OSS has continued to grow, estimates of its market penetration have in different areas: operating systems, servers, and specialized languages. The most widely used operating system on the internet globally is the Linux-based Android operating system (GlobalStats statcounter 2018). As of July 2018, Apache is the most frequently used server on the internet (W3Techs, 2018). Based on the cost of the nearest available substitute, Greenstein and Nagle (2013) estimated the value of capital stock of Apache software in use in 2013 at between $512 million and $12.8 billion.

OSS is increasingly developed and used by businesses; the top institutional contributors to the code repository Github in 2017 are Microsoft and Google. Nonbusiness contributors are

active as well on Github, with Stanford University, the Massachusetts Institute of Technology, The Broad Institute, and UC Berkeley among the top contributors. Additionally, contributors from the US Federal Government include Sandia National Laboratory and General Services Administration (Hoffa 2017). Recent policies of the U.S. Federal government now promote the posting and sharing of software source code developed by or for the Federal government (Scott and Rung. 2016).

While this particular policy to promote reuse and sharing of software created with public funding is relatively new, public funding has had an important and as yet not fully accounted role in the creation of open source software (OSS). How much of the OSS currently downloaded and used has been created with public funding, in the U.S. or in other nations? We currently don't know. We do know that many of the most widely used OSS tools available have been developed in universities and other publicly-funded institutions. Apache is open source server software developed with federal and state funds at the National Center for Supercomputing Applications in Illinois. Linux was developed at the University of Finland. The R language was developed at the University of Auckland in New Zealand by professors for use in their teaching laboratory, with extended development by Hadley Wickham at Iowa State University (Ihaka, 1998; XXX). Independent nonprofit institutions have also played an important role in OSS, assuming the management and coordination of OSS projects; examples include the Apache Software Foundation, the Mozilla Foundation, and the Linux Foundation.

The work of OSS developers across sectors of the economy also highlights the role of non-business activities that result in important innovations. As the OECD's innovation measurement manual defines it: "the implementation of a new or significantly improved product (good or service) or process, a new marketing method, or a new organization method in business

practices, workplace organization, or external relations" (OECD and Eurostat, 2005). The creation of new software products and their significant improvement through version upgrades and package extensions that are systematically shared with potential users fit well within this definition. As policy-makers seek better information to identify innovations and the systems that create them, OSS provides a clear window into a previously underexplored area: innovation in all sectors of the economy (Gault, 2018).

The contribution of this paper is two-fold. First, we present a framework for the presentation of statistics about OSS, categorizing it as a subcomponent of a recognized investment category in national economic accounts: own-account software. Second, we present and implement a prototype methodology for resource cost estimation for OSS investment that is conceptually consistent current measurement of own account investment. Our "bottom-up" methodology uses data collected from within OSS package code and data about OSS packages; we estimate that the R programs currently available to users have a cumulative resource cost in the magnitude of XXX to XXX, equivalent to XXX. We also show how these same data can be used to measure the relative impact of OSS packages, showing the linkages between packages that magnify impact.

Literature Review

A decade ago interest in OSS centered around understanding the motivations for participation in open source software (OSS) projects and evaluations of its growth potential. Surveys conducted in the early 2000s describe the contributor community (Ghosh, et al. 2002; Boston Consulting Group 2002; David, Waterman, Arora 2003). While the motivations described for participation include skill development, creativity, and interest in the open source community, user-need and functionality consistently rate as critically important. A Boston

Consulting Group survey (2002) probed the motivations for participating in OSS project; over 60% reporting either work or non-work functionality.

For many academics and researchers, software tools and data bases are by-products of their own work that can also be used by other academics as well (Gambardella and Hall 2005). Advantages of OSS include the ability to scale customization projects and to resolve program bugs quickly through many users (Lerner and Tirole. 2005). OSS communities can also be viewed as user innovation networks, where contributors more successfully develop solutions to their own software needs through the OSS community (von Hippel 2005).

Interest in better measurement of the economic impact of computer software and the increased digitization of knowledge led to parallel development in national economic accounting in many countries. For example, Gross Domestic Product (GDP) statistics for the US have treated computer software as investment since 1999, extending this treatment to research and development expenditures and entertainment and literary originals in 2013 (BEA 2013). Beyond these three categories, Corrado, Hulten, and Sichel (2005) provide a framework for consistent accounting for expenditures on intangibles that generate future benefits. Arguing that public expenditures yielding long-lived returns should be understood as investment, Corrado, Haskell, and Jona-Lasinio (2015) propose a public investment category, information, scientific, and cultural assets, which includes software and databases along with R&D mineral exploration and cultural products. They argue that better accounting of public investment in intangibles would provide a more complete picture of economic growth (CSL, 2015).

Figure 1 shows software investment in the US for 2016 (in current dollars) from the Bureau of Economic Analysis. Overall, software accounts for more than $400 billion dollars a year in fixed investment in 2016. Less than $100 billion of this was for public sector investment

(federal labs and facilities, public universities, state and local government entities). Most, more than $300 billion dollars, is for private sector investment in software. For comparison, BEA reports investment in R&D on the same set of tables. Private sector R&D investment is $320 billion and the public sector (government)'s investment was $170 billion in 2016.[1]

Software as investment has three types, based on data source. These are prepackaged, custom, and own account. Prepackaged and custom software are purchased inputs, and industry receipts data are used for these estimates. Own account software is not purchased or sold. For measurement purposes, 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 and Grimm 2000).

From a classification perspective, this category of own account software is where OSS belongs. OSS fits the definition of own account software as laid out in BEA (2000) and OECD (2010) methodologies: OSS is original software that provides benefits by using the software directly, and these benefits are expected to last for years. The key to its treatment as intangible investment is that resources shift from consumption to creating something that provides value in use over a long period of time to its creator.

Open source software is a component of own account software. We expect to find measurable amounts in universities, both private and public, in institutions run by state and local governments, and an in Federal investment. Private university-created OSS would be within the private estimates, as would any created by private nonprofit institutions.

[1] This paper was prepared prior to the release of BEA's comprehensive revision on July 27. The treatment was changed for R&D expenditures that overlap with own-account software expenditures, and thus data revised. [https://blog.bea.gov/2018/07/03/coming-july-27-gdp-statistics-updated-to-keep-pace-with-ever-changing-economy/]

Definitions

Innovation: The implementation of a new or significantly improved product (good or service) or process, a new marketing method, or a new organization method in business practices, workplace organization, or external relations. This definition is based on the Oslo Manual, which provides guidance for internationally comparable measurement (OECD and Eurostat 2005).

Intangible capital: As a subcategory of capital investment, intangible capital includes economically useful ideas, creations, or plans that can be reused repeatedly in production over a long period of time. Research and development expenditures, computer software and databases, and entertainment and literary originals are counted as intangible capital in national economic accounts. Additional categories of intangibles that have similar features are economic competencies, made up of brand equity, firm-specific human capital, and resources devoted to organizational development (Corrado, Hulten, and Sichel,2005).

Open Source Software: computer software with its source code made available with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose. For this paper, we treat as open source any software language or package with an Open Source Initiative (OSI)-approved license.

Own account software: A category of software investment in national economic accounts. Own-account software is long-lasting software created for internal use, rather than as a market product.

Production Ready Release: A software package that is ready for production in its current ecosystem

Registry: A location that hosts, manages, and distributes OSS. The Comprehensive R Archive Network (CRAN) is an example.

Repository: A location that hosts and maintains different versions of software programs. Github is an example.

Empirical Framework

Robust indicators of the creation and use of OSS can provide more complete measures than currently available on the outputs and impacts of public investments in Science and Engineering. The paper describes two parallel measurement tracks, one focused on impacts and one focused on creation cost.

Counts and Linkages: The production and delivery of open source software through publicly accessible websites provide harvestable count and linkage data for software languages and packages. These data can be analyzed with methods developed for bibliometrics and patent analysis. Since their development by Derek Price (needs citation XXX) through peer-reviewed publication citation indexes, indicators of research activity and impact calculated from databases covering scientific article publications and citations are frequently used measures of research intensity and impact. Many well-developed methodologies and extensions exist, and a research community continues to publish. For patents, Hall, Jaffe, and Trajtenberg (2001) released a comparable electronic data set for U.S. Patent and Trademark Office (USPTO) patents along with methodological information for its use in the NBER patent citation file. Articles and patents share features of skewed distributions, where only a small number of articles and patents generate measurable impact, and those that have impact can be blockbusters.

Cost of production: Investment in own-account software, created for internal use, is estimated for national investment statistics based on its cost of production. This approach is the same one used for other types of products not sold in the market. BEA researchers have proposed a similar methodology to measure advertising-supported products: when bundled with advertising, free content created in the business sector can be valued based on its production cost (Nakamura, Samuels, and Soloveichik (2018).

Production costs for own account software include those for analysis, design, programming and testing, and exclude maintenance and repair (Parker and Grimm, 2000). As originally described, cost of production is the sum of: labor costs and intermediate inputs (such as materials and supplies and overhead). In US economic statistics these costs are estimated with the mean wage rate for computer programmers and system analysts, adjusted for compensation costs, and multiplied by the number of these workers in each industry. Wage, employment and compensation are from the U.S. Bureau of Labor Statistics data. To account for time spent on tasks other than software development, BEA uses an adjustment ratio from a survey of software developers time spent on different tasks of 487 businesses (Boehm,1981). Non-labor costs for OSS development are estimated with industry production ratios (Parker and Grimm 2000).

To illustrate the contribution to OSS from public spending, more detailed breakdown by sector is needed, both for relative measures based on counts and uses, or a dollar-denominated cost measures. Figure 2 shows the categories needed to provide a cross-walk between existing published data on software investment and the level of detail needed to show where open source software is created.

Data Collection and Preparation

Keller et al (2018) describes the overall approach to exploring the use of alternative data sources and non-survey data to improve and extend indicators of science and engineering activity and of innovation using alternative data sources. This approach includes structured processes to discover, acquire, profile, clean, link, explore the fitness-for-use, and statistically analyze the data. To explore the potential to develop indicators of open source software, we exploit publicly available metadata about individual packages and their contributors, as well information within