Measuring the Cost and Impact of Open Source Software as Intangible Capital

J. Bayoán Santiago Calderón¹  Brandon L. Kramer¹
Gizem Korkmaz¹  Carol A. Robbins²
Aaron D. Schroeder¹  Sallie A. Keller¹

¹BIOCOMPLEXITY INSTITUTE

2020 FCSM CSPOS Webinar on Blended Data
Project Background

- Cooperative Agreement between UVA and NCSES\(^1,2\)
- Exploration and application of data science research for the enhancement, improvement, or redesign of current surveys on R&D, Innovation, S&E education and S&E workforce
- The multi-year project aims to improve indicators of research output and innovation activities, including developing reliable data on the creation of open source software throughout the economy

---

\(^1\)This work is supported by U.S. Department of Agriculture (58-3AEU-7-0074)

\(^2\)The views expressed in this work are those of the authors and not necessarily those of their respective institutions.
Presentation Overview

- Part 1: Measuring Scope and Impact of OSS
  - Project overview
  - Defining the universe of OSS
  - Cost estimation
  - Proposed impact measures

- Part 2: International OSS collaboration
  - Contributor networks
  - Country-to-country networks
Introduction

Current NCSES and other economic indicators do not measure the scope and impact of open source software developed outside the business sector.\(^3\)

- How much open source software is in use? (stock measure)
- How much is created each year? (flow measure)
- What sources of data and variables could be used to measure cost and impact?

---

What is Open-Source Software?

- Software that is published under an Open Source Initiative (OSI) approved license
- OSI-approved licenses establish permissions (e.g., use, inspect, modify, distribute, attribution) and limitations (e.g., liability, warranty).
- Most common licenses are: MIT, Apache, GPL, etc.
Project vs. Repository

“An open source project is not the collection of lines of code that make up its repository. At its core, an open source project is the sum of commitment and sacrifice of time, energy, money and sanity by the contributors to such project. By such contributions, I don’t mean contributions in code, or contributions to (in our case) JuliaLang/Julia, but any such contributions to the community whether in code, in documentation or in community engagement (evangelism, organizing meetups, etc.).”

- Keno Fisher (Co-Founder/CTO of Julia Computing)
Defining the Universe of Open Source Software

- **Guiding Questions**
  - Package vs template vs homework vs Stack Overflow answer
  - Is it usable (i.e., production-ready, maintained)?
  - Is it discoverable (and installable)?

- **Two proposed solutions:**
  1. **Packages for programming languages**
     These are published codebases that are discoverable and installable through a registry and package manager.
  2. **GitHub repositories**
     Repositories on GitHub, the world’s largest remote hosting platform for Git version control.

---

Approach 1: Packages

**Data collection strategy**

<table>
<thead>
<tr>
<th>OSS Definition</th>
<th>Registry</th>
<th>Repository (GitHub)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Open Source Initiative (OSI) -Approved License</td>
<td>- Language-specific package managers (e.g., CRAN, PyPI)</td>
<td>- Commit Data (who, what, when)</td>
</tr>
<tr>
<td>- Production Ready</td>
<td>- Contains metadata</td>
<td>- License</td>
</tr>
<tr>
<td>- Release for Current Ecosystem</td>
<td>- Continuous integration</td>
<td>- Profile of contributors</td>
</tr>
</tbody>
</table>

- The registry data was collected using web harvest techniques.
- All CRAN and PyPI data as of July 2017.
Package Census: How much OSS is available?

<table>
<thead>
<tr>
<th>Language</th>
<th>R</th>
<th>Python</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registry</td>
<td>CRAN</td>
<td>PyPI</td>
</tr>
<tr>
<td>Number of packages</td>
<td>13,719</td>
<td>164,836</td>
</tr>
<tr>
<td>OSI-approved &amp; production ready</td>
<td>13,143</td>
<td>15,043</td>
</tr>
<tr>
<td>Packages on GitHub</td>
<td>4,407</td>
<td>11,016</td>
</tr>
<tr>
<td>Packages on GitHub (analysis)</td>
<td>4,358</td>
<td>9,773</td>
</tr>
</tbody>
</table>
Approach 2: GitHub Repositories

Data collection strategy

- Find public repositories with an OSI-approved license
- Collect development activity information (e.g., commits, additions)
- Collect information about contributors (e.g., organization, location)
- Supplement the GHTorrent Project\(^5\) that collects public scope data using the GitHub v3 RESTful API, with additional variables collected using the GitHub v4 GraphQL API.

---

GitHub Census: How much OSS is available?

The GitHub dataset is comprised of around **3.3M distinct contributors** and **7.8M repositories**.
Most repositories have fewer than 5 contributors.
Cost Estimation of OSS

Constructive cost model (COCOMO II\(^6\)) developed in software engineering

1. Kilo-lines of code (KLoC) represent effort
   - Effort is a nonlinear function of complexity and lines of code
     \[
     \text{Effort} = 2.4 \times (\text{KLoC})^{1.05}
     \]
   - Nominal development time = \(2.5 \times \text{Effort}^{0.38}\)
   - Development cost = Monthly wage \(\times\) Nominal development time

2. Estimate resource cost with wage equivalent for 2017
   - Computer programmers, software developers

3. Estimate non-wage costs adapting OECD and BEA methods

---

Cost Estimation

- Cost estimates for open-source packages registered in PyPI and CRAN packages hosted in GitHub: **$2.4 Billion** (14,131 total packages)
- Cost estimates based on GitHub public machine detectable OSI-approved licensed repositories: **$928 Billion**

<table>
<thead>
<tr>
<th>Repository</th>
<th>KLoC Added</th>
<th>Cost in Million $</th>
</tr>
</thead>
<tbody>
<tr>
<td>ringmesh/RINGMesh</td>
<td>1,124,359</td>
<td>18.1</td>
</tr>
<tr>
<td>sk-/python2.7-type-annotator</td>
<td>350,656</td>
<td>11.3</td>
</tr>
<tr>
<td>LuxCoreRender/LuxCore</td>
<td>298,708</td>
<td>10.6</td>
</tr>
<tr>
<td><strong>Total for 7.8M repos</strong></td>
<td><strong>471M</strong></td>
<td><strong>928K</strong></td>
</tr>
</tbody>
</table>

GitHub data collected up to 2019.

Table: Top 5 GitHub repositories with the highest KLoC and cost.

---

7The cost calculations were based on data collected through July 2019 in 2017 dollars. We are currently updating the cost estimates.
**Impact of OSS: Downloads and Dependency Networks**

<table>
<thead>
<tr>
<th>Package</th>
<th>2018 Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rcpp</td>
<td>3,519,510</td>
</tr>
<tr>
<td>rlang</td>
<td>2,893,889</td>
</tr>
<tr>
<td>stringi</td>
<td>2,610,184</td>
</tr>
<tr>
<td>stringr</td>
<td>2,511,011</td>
</tr>
<tr>
<td>ggplot2</td>
<td>2,495,315</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Package</th>
<th>Reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>ggplot2</td>
<td>105,774</td>
</tr>
<tr>
<td>plyr</td>
<td>101,596</td>
</tr>
<tr>
<td>digest</td>
<td>99,774</td>
</tr>
<tr>
<td>stringr</td>
<td>98,086</td>
</tr>
<tr>
<td>colorspace</td>
<td>93,590</td>
</tr>
</tbody>
</table>

---


International OSS Collaboration

- International collaboration doubled in academic papers since 1990
- More governmental funding for projects that developed through international collaboration tend to lead to higher impact publications
- Understanding international contributions in the context of OSS will help explain:
  - Which countries are most likely to contribute to OSS?
  - Which countries are most likely to collaborate across international boundaries?
  - What is the structure of international collaborations and how does it change over time?
  - Which are the most influential countries in the OSS ecosystem?
Social Network Analysis

- A scientific framework that helps researchers conduct structural analysis on various types of relational entities, including contributors and countries, that are visualized as nodes and edges in graphs.

- We will develop and analyze two types of networks:
  - GitHub contributor networks
  - Country-to-country collaboration networks
Our original dataset is comprised of 3.3M distinct contributors and 7.8M distinct repositories along with GHTorrent’s location data. To examine international collaborations, we reduced the dataset to only include logins with valid country codes, which included 700K contributors and 3.6M repositories dating from 2008 to 2019.

**Total Contributors for Top-10 Countries (GitHub, 2008-2019)**

- USA: 200
- China: 50
- UK: 50
- Germany: 50
- India: 50
- Canada: 50
- France: 50
- Brazil: 50
- Russia: 50
- Japan: 50
Method

To convert these data into network format, we “projected” a bipartite network of login and repository information into single-mode contributor and country networks.
Method

- The contributor network is comprised of
  - Nodes represent contributors
  - Edges correspond common repositories that users contribute

- The country-country network is comprised of
  - Nodes represent countries
  - Edges correspond to international collaborations

- Analyzed networks using R’s igraph and ggraph and Gephi
Contributor Network Dynamics

From 2008-2019, the number of nodes, edges, and overall commits grow exponentially in the GitHub contributor networks.
Country-Country Networks

<table>
<thead>
<tr>
<th>Network Descriptives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>214</td>
</tr>
<tr>
<td>Edges</td>
<td>11K</td>
</tr>
<tr>
<td>Density</td>
<td>0.48</td>
</tr>
<tr>
<td>Avg Deg Cent</td>
<td>103</td>
</tr>
<tr>
<td>Avg Btw Cent</td>
<td>57.5</td>
</tr>
</tbody>
</table>
International OSS Collaboration

Degree vs. Betweenness Centrality

Centrality measures provide insights about relative influence of countries (e.g., the US has a higher capacity to bridge different countries)
Degree vs. Betweenness Centrality
Community Detection

We used Newman’s modularity algorithm to examine country groupings based on network properties.

Modularity Groupings for Country-Country Collaboration Networks (GitHub, 2008-2019)

Note: Groupings Calculated with Loops in Network
Conclusions

- **GitHub Contributor Networks**
  
  US-based users are most likely to contribute to GitHub

  Over the past decade, GitHub’s network size has grown exponentially with a similar structure to scientific collaboration networks

- **Country-Country Networks**
  
  Centrality measures reveal that some countries (like the USA) have more influence in the OSS ecosystem

  The country-country network clusters into four communities, reflecting a combination of socio-political factors that shape OSS development
Policy Implications

- **Richer data for national accounts**
  Currently, the national accounts captures software through three categories: (1) pre-packaged, (2) own-account, and (3) custom software which significantly exclude OSS.

- **Use of administrative data to complement survey research**
  Our approach of re-purposing administrative data, especially of unconventional data such as repositories opens a new venue for statistical agencies to address current data limitations.

- **Use of network analysis**
  Helps to examine the relational structure of OSS collaborations and how those connections may relate to the global economy.
Questions?