Measuring the Scope and Impact of Open Source Software

Government Advancement of Statistical Programming (GASP) Workshop 20

23 September 2019

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE

NSF NCSES National Center for Science and Engineering Statistics

The NCSES and UVA
Team Open Source Software (TOSS)

NCSES	Carol Robbins	Senior Analyst
UVA	Gizem Korkmaz	Research Associate Professor
UVA	Aaron Schroeder	Research Associate Professor
UVA	Bayoán Santiago Calderón	Postdoctoral Research Associate
UVA	Brandon Kramer	Postdoctoral Research Associate

Open Source Software (OSS)

“A 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.” (Open Source Initiative)

It is developed *within and outside* of the private sector

- *universities* (e.g., Stanford, MIT, UC Berkeley),
- *businesses* (e.g., Microsoft, Google),
- *government research institutions* (e.g., Sandia National Lab),
- *nonprofits*, and
- *individuals*

Current NCSES and other economic indicators do not measure the *value of open source software* outside the business sector.

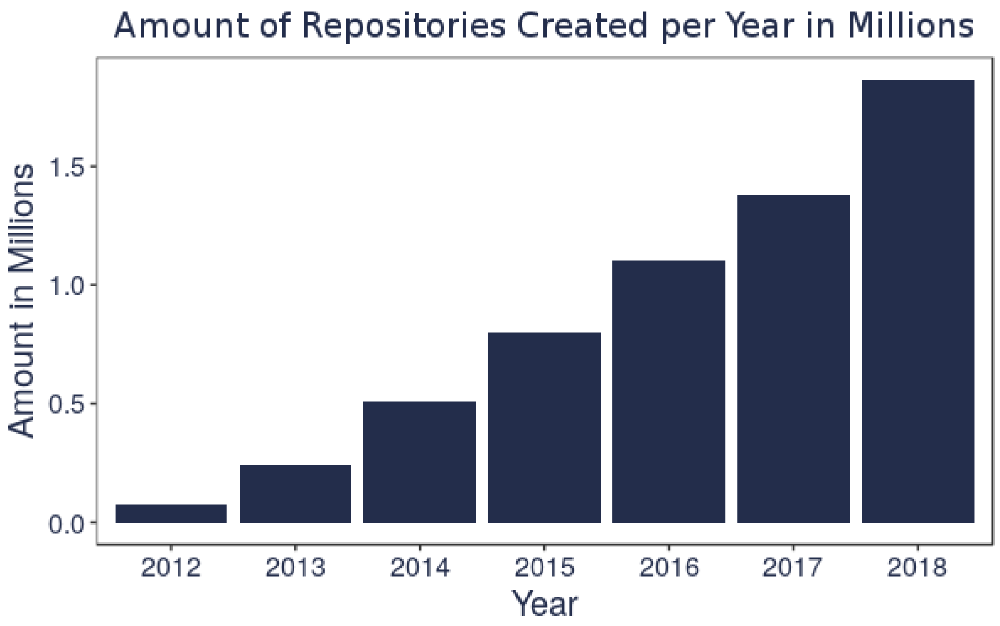
Research Questions

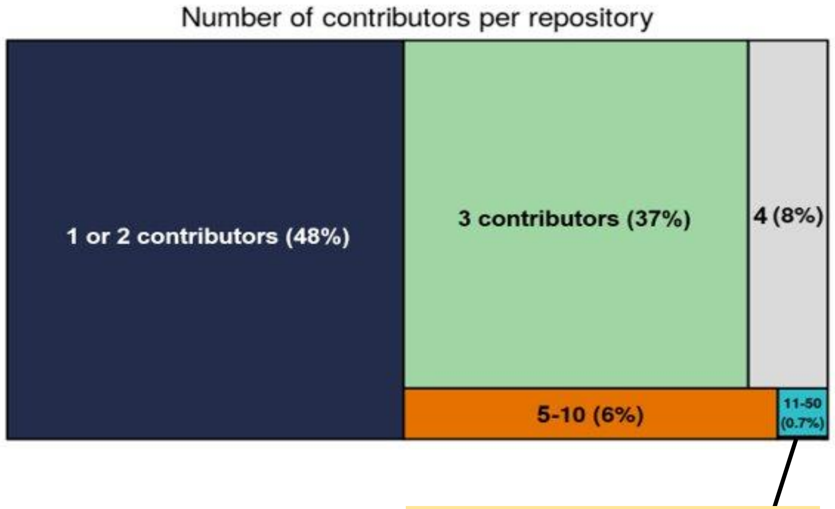
- How much open source software is in use? (*stock measure*)
- How much is created each year? (*flow measure*)
- What types can be identified?
- Who creates it? (*Sectors:* Business, Government, Academia, Households, Nonprofits, Foreign)
- What *sources of data and variables* could serve as proxies for measuring attributes and economic value of open source software?
  - Counts of OSS, by sector (e.g., university, government), by technology
  - Attributes, e.g., # citations, #downloaded,

OSS Universe: Projects on GitHub

- **55.1M** repositories with commits 2012-2018
- **7M** repositories with OSI-approved licenses on GitHub as of July 2019.

  Of those, we analyzed **4.9M** repos that have at least one commit.
- There are **2.8M** unique OSS contributors

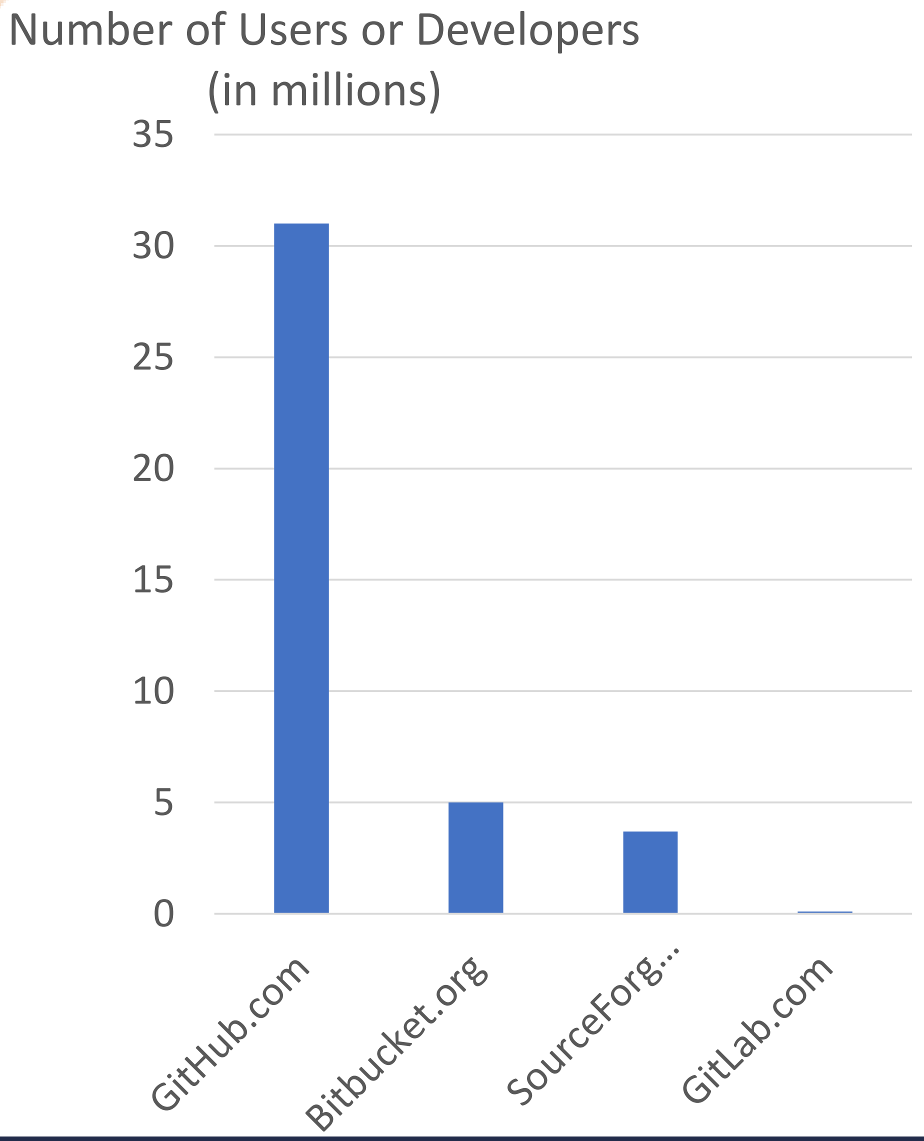

Amount of Repositories Created per Year in Millions


Number of contributors per repository

> 50 contributors (0.3%)

Most repositories have fewer than 5 contributors.

OSS Universe: Programming Languages

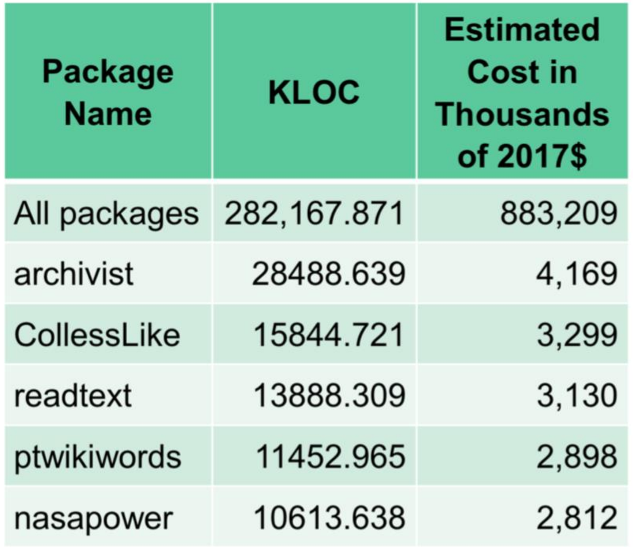


Language	R	Python
Package manager	CRAN	PyPI
Number of packages	13,719	164,836
OSI-approved & production ready	13,143	15,043
Packages on GitHub	4,407	11,016
Packages on GitHub (analysis)	4,358	9,773

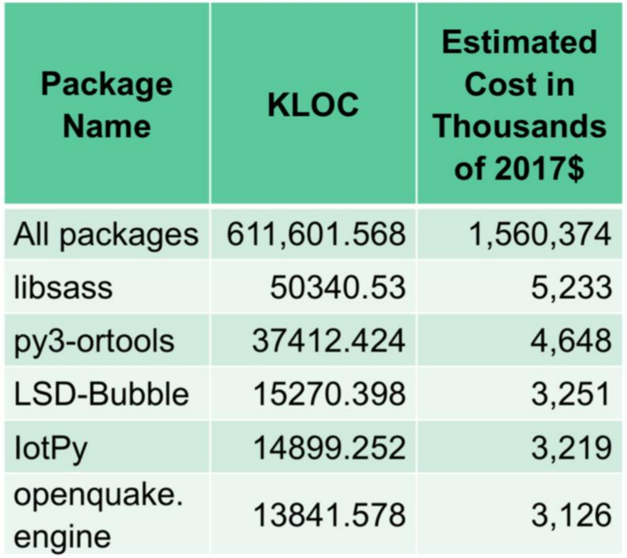
Value of OSS: Cost of Software Package Creation

- Identify number of people involved each package’s development
- Estimate time spent on software development using **Kilo-lines of code (KLOC)**
- Estimate **resource cost** with wage equivalent for 2017
  - Using average compensation for **computer programmers**
  - Occupation Employment Survey, Bureau of Labor Statistics
- Estimate **non-wage costs**
  - Adapting BEA (Bureau Economic Analysis) and OECD (Organisation for Economic Co-operation and Development) methodologies

Cost of R and Python Packages developed on GitHub



Package Name	KLOC	Estimated Cost in Thousands of 2017$
All packages	282,167.871	883,209
archivist	28488.639	4,169
CollessLike	15844.721	3,299
readtext	13888.309	3,130
ptwikiwords	11452.965	2,898
nasapower	10613.638	2,812

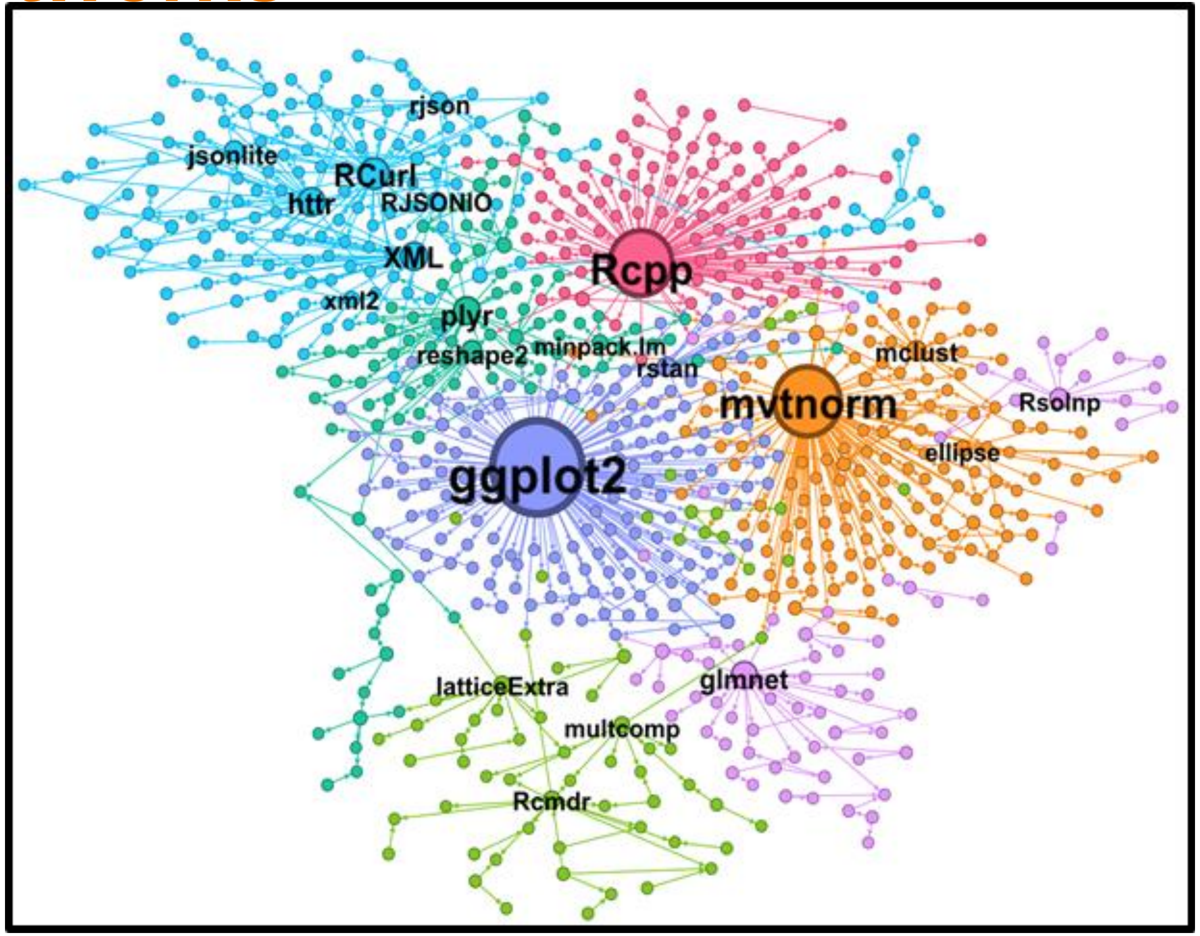


Package Name	KLOC	Estimated Cost in Thousands of 2017$
All packages	611,601.568	1,560,374
libsass	50340.53	5,233
py3-ortools	37412.424	4,648
LSD-Bubble	15270.398	3,251
IotPy	14899.252	3,219
openquake. engine	13841.578	3,126

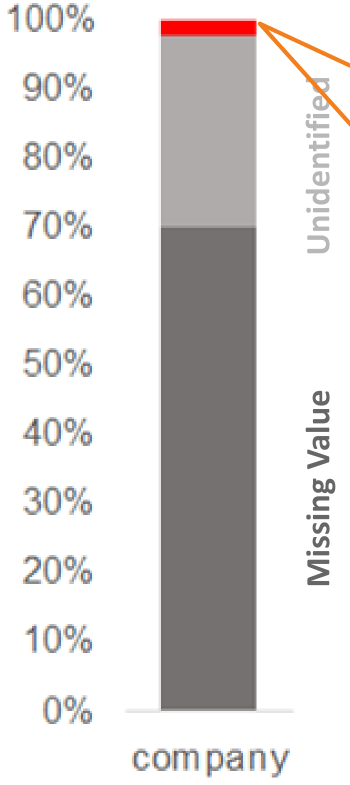
Impact of OSS: Downloads and Dependency Networks

Package	2018 Downloads
Rcpp	3,519,510
rlang	2,893,889
stringi	2,610,184
stringr	2,511,011
ggplot2	2,495,315

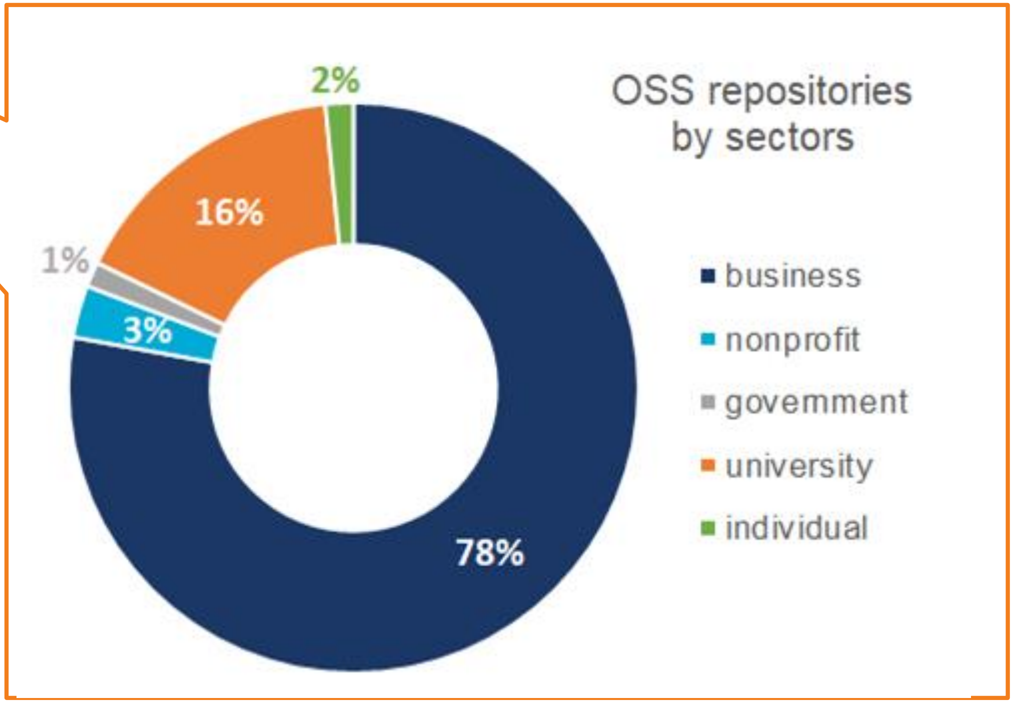
Package	Reuse
ggplot2	105,774
plyr	101,596
digest	99,774
stringr	98,086
colorspace	93,590



Limitations: Sectors of OSS



In the data, only ~2% of the users' sectors can be identified



- We use the self-reported company field in contributors profile.
- Only 2% of the contributors can be identified.
- OSS is becoming more permissive as businesses contribute more code.

Next Steps

- Get more detailed data on the OSS repositories, including additions and deletions to estimate the development cost using the lines of code
- Obtain contributor emails to improve the sector analysis
- Conduct network analysis to study interactions between contributors and OSS projects, and diffusion of OSS innovation.

Products

Peer-Reviewed Publications

Keller, S.A.; G. Korkmaz; C. Robbins; and S. Shipp. 2018. “Opportunities to Observe and Measure Intangible Inputs to Innovation: Definitions, operationalization, and examples.” *Proceedings of the National Academy of Sciences (PNAS)*. 115 (50), 12638-12645.

Korkmaz, G.; C. Kelling; C. Robbins; and S. Keller. 2018. “Modeling the Impact of R Packages Using Dependency and Contributor Networks.” *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining* (ASONAM), pp. 511-514. IEEE.

Conference Presentations/Papers

- National Bureau of Economic Research (NBER) Conference on Research in Income and Wealth (CRIW) Conference: Big Data for 21st Century Economic Statistics, Bethesda, MD, Mar. 2019.
- 2019 Women in Data Science Conference, Charlottesville, VA, Mar. 2019.
- International Association for Research in Income and Wealth (IARIW) 35th General Conference: Innovation and the Digital Economy, Copenhagen, Denmark, Aug. 2018.
- IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM), Barcelona, Spain, Aug. 2018.
- International Monetary Fund (IMF) Statistical Forum on Measuring Economic Welfare in the Digital Age: What and How? Washington D.C., Nov. 2018.
- NBER Conference on Research in Income and Wealth (CRIW) Pre-Conference: Big Data for 21st Century Economic Statistics, Cambridge, MA, Jul. 2018.
- Interagency Council on Statistical Policy (ICSP) Big Data Day, Committee on National Statistics (CNSTAT), Washington, DC, May 2018.
- Federal Committee on Statistical Methodology (FCSM) 2018 Research and Policy Conference, Washington DC, Mar. 2018.
- Arthur M. Sackler Colloquia of Sciences: Modeling and Visualizing Science and Technology Developments, Irvine, CA, Dec