Open Source Software as Intangible Capital: Measuring the Cost and Impact of Free Digital Tools

Carol A. Robbins[1], Gizem Korkmaz[2], José Bayoán Santiago Calderón[3], Daniel Chen[2], Claire Kelling[4] , Stephanie Shipp[2], Sallie Keller[2]

Measuring Economic Welfare in the Digital Age: What and How?
IMF Statistical Forum
Washington, D.C.
November 19–20, 2018
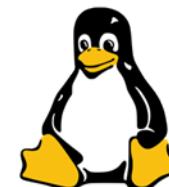
1) National Center for Science and Engineering Statistics, National Science Foundation ; 2) Social & Decision Analytics Division, Biocomplexity Institute & Initiative, University of Virginia; 3) Claremont Graduate University; 4) Pennsylvania State University

The views expressed in this paper are those of the authors and not necessarily those of their respective institutions.

Why Open Source Software?

Long-lasting Benefits, Zero Purchase Cost, Creates New Products

- We access detailed data about these new products and tools from repositories

  - OSS shared on Code.gov and Github

- Develop framework for measuring cost and analyzing contributors

  - Estimate resource cost for four OSS Languages

- Explore network relationships between OSS packages as measures of impact

Open Source Projects by Federal Government Organization
Top 5 by number of projects
for projects started before January 1, 2018

Organization Name	Total Projects on Code.gov	Number of Projects Linked to Github collection	Kilo-lines of code (kloc)	Commits	Number of contributors
Total	4,457	2,688	2,486,210	950,625	8,292
General Services Administration	1,501	1,368	266,860	318,676	4,631
Department of Energy	899	704	1,219,835	485,726	2,433
Consumer Financial Protection Bureau	261	243	753,447	49,781	334
National Aeronautics and Space Administration	998	141	179,917	51,936	358
Environmental Protection Agency	156	61	14,327	4,711	78

Note: for projects started before January 1, 2018
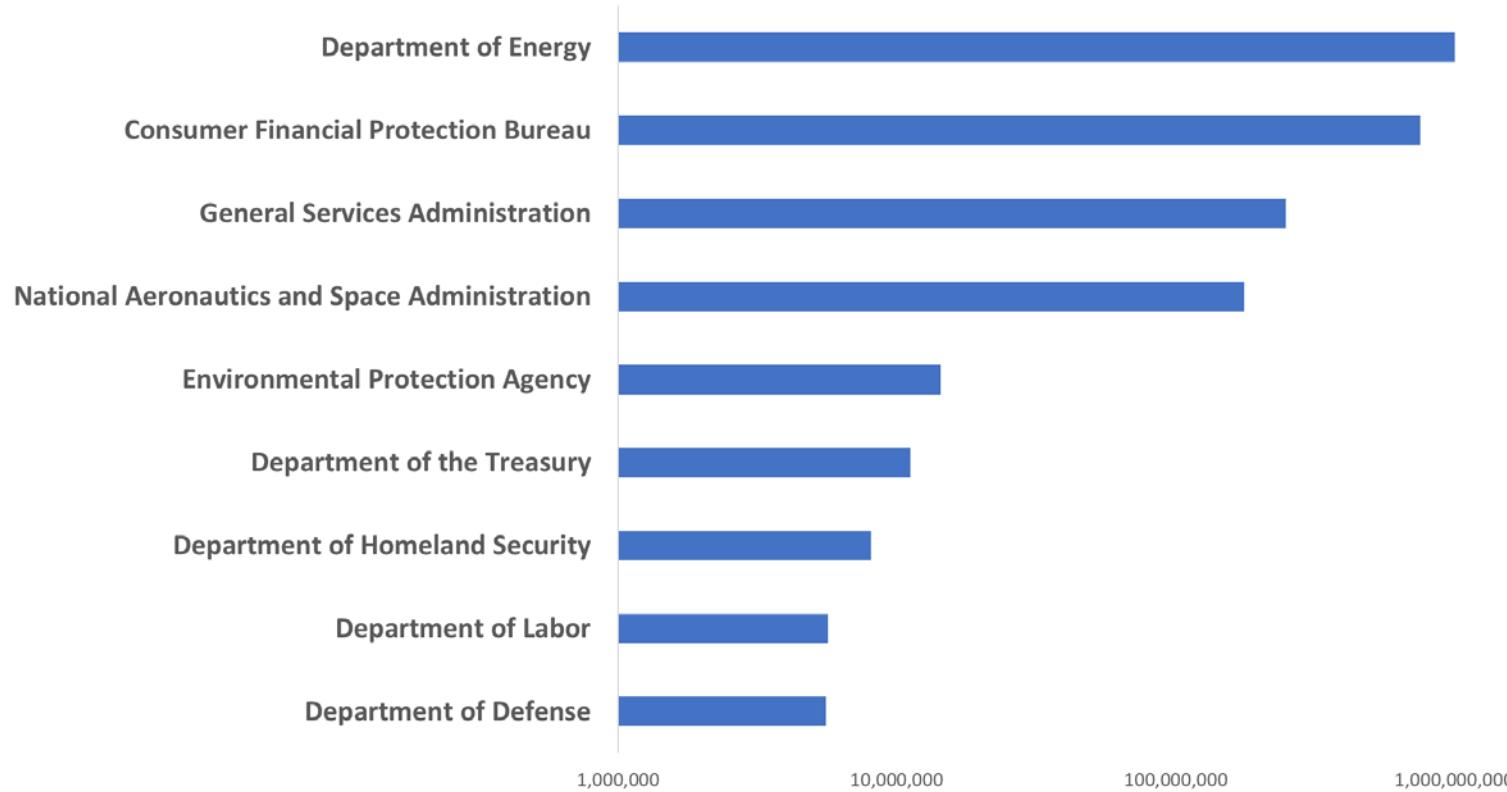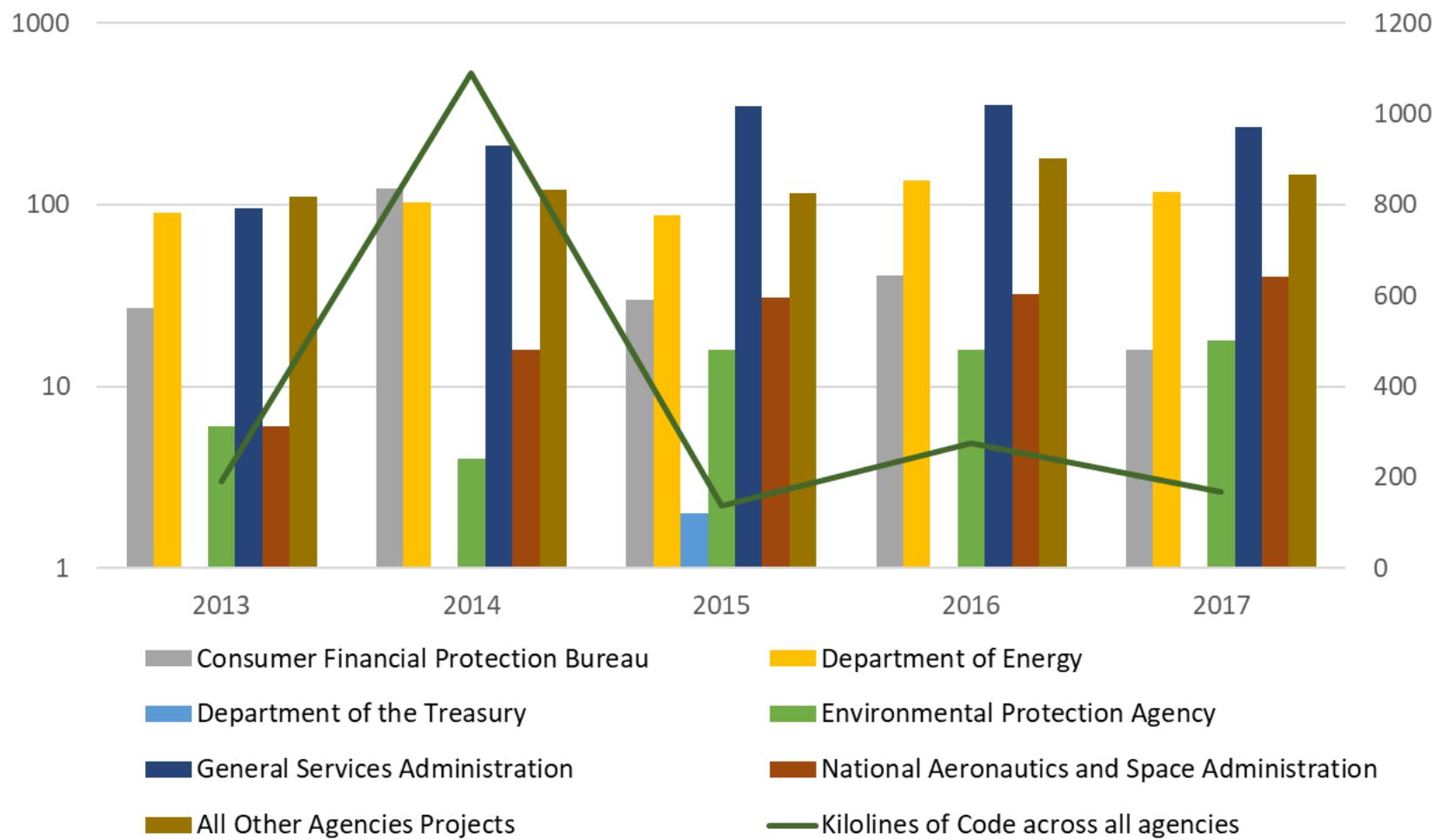
Federal Contributions to Code.go are Widespread, but Vary Greatly Across Agencies
Agencies Contributing more that 1,000,000 lines of Code
Lines of Code in Projects Started between 2008 to 2017, log scale

Note: Calculated from the projects on code.gov that are also posted on github.

Most Projects on code.gov come from a Handful of Agencies
projects and klines of code, 2013-2018

Across all Agencies, klines of code

Legend:
- Consumer Financial Protection Bureau
- Department of Energy
- Department of the Treasury
- Environmental Protection Agency
- General Services Administration
- National Aeronautics and Space Administration
- All Other Agencies Projects
- Kilolines of Code across all agencies

Note: Calculated from the projects on code.gov that are also posted on github.

Framework to Highlight Investment in OSS Software by Different US Sectors

Software subcategory of Intellectual Property Products Investment	Private Sector			Public Sector			Household Sector	Rest of World
	Business	Other private nonprofits	Higher education	Higher education	Federal Government and FFRDCs	Non-federal government, ex. Higher Ed.		
Prepackaged								
Custom								
Proprietary								
Open Source (OSS)								
Own-account								
Proprietary								
Open Source (OSS)								

OSS Languages

Data Collection				
Language	Package manager	Number of packages	OSI-approved & production ready	Packages on Github
R	CRAN	12,614	11,886	3,396
Python	PyPi	143,047	7,392	3,804
Julia	Pkg.jl	2,040	1,324	1,324
JavaScript	CDNJS	3,367	3,367	3,213
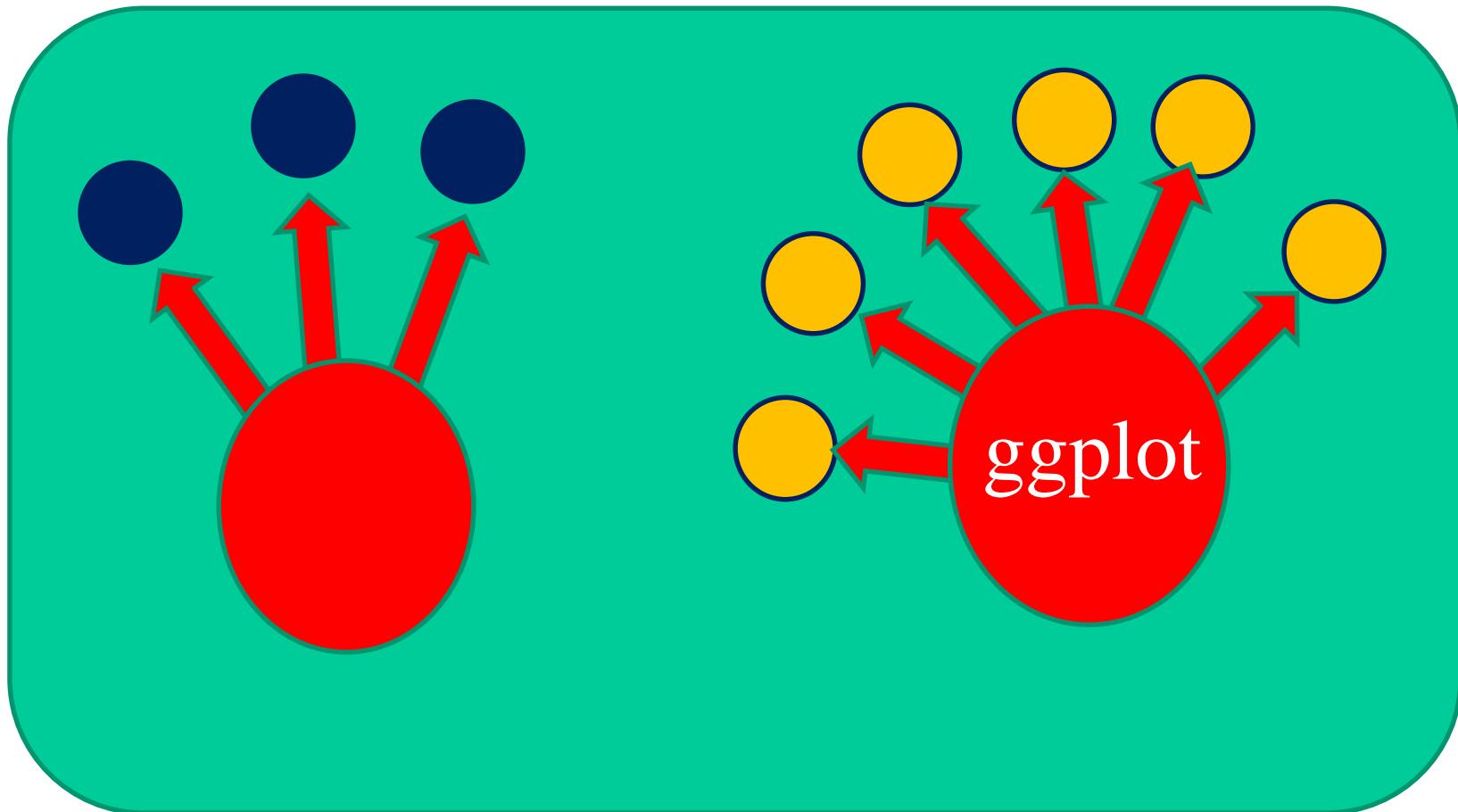
# Cost of OSS Software Package Creation
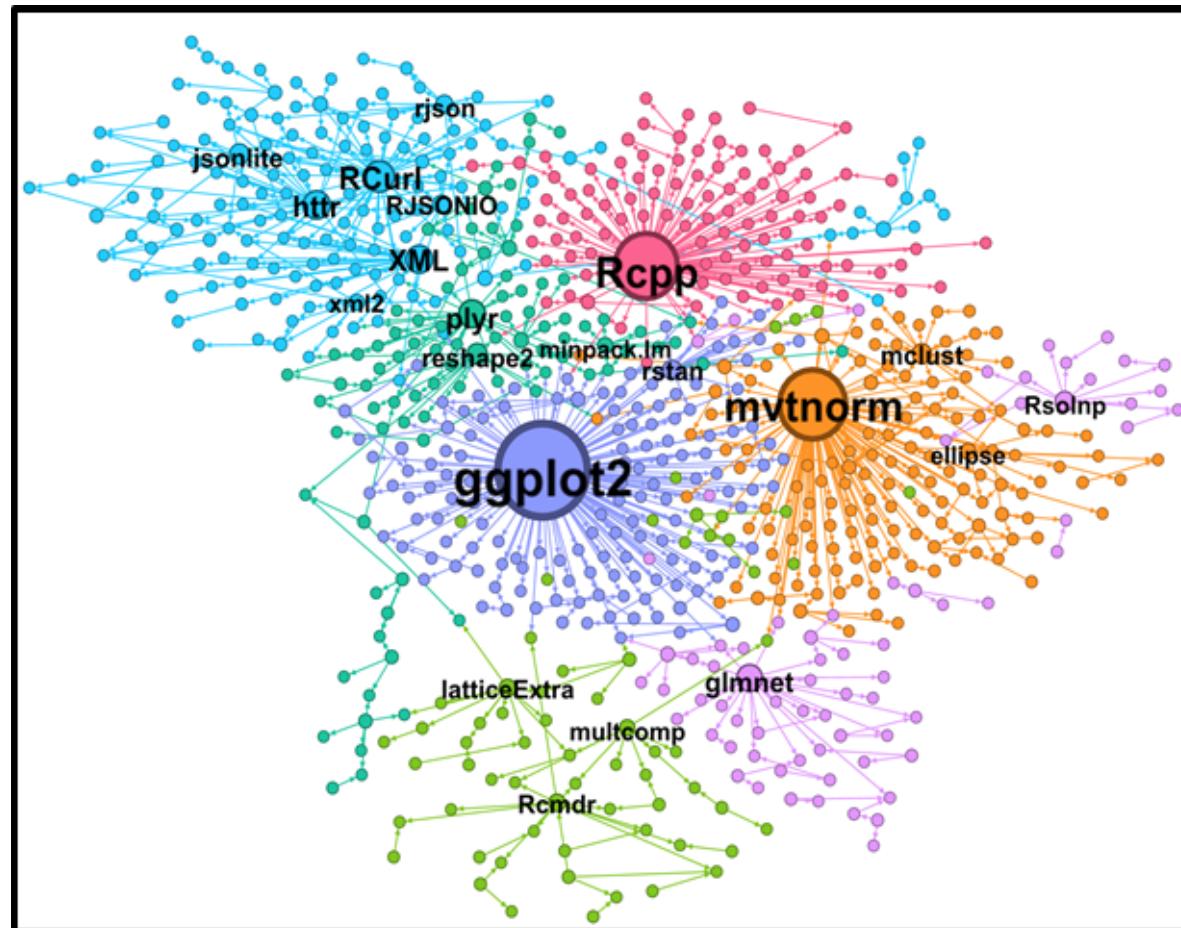
1) Kilo-lines of code represent effort
   - Effort is a function of complexity and lines of code
   - historical software project factors

2) Estimate resource cost with wage equivalent for 2017
   - Computer programmers, software developers
   - Occupation Employment Survey, Bureau of Labor Statistics

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

Top 5 packages with the highest total cost (in USD)							
CRAN (R)		**PyPi (Python)**		**Julia (Julia)**		**CDNJS (JavaScript)**	
Package	Cost	Package	Cost	Package	Cost	Package	Cost
googleAnalyticsR	4.4M	Nupic	3.9M	GeoStatsImages	4.2M	Webkit.js	7.6M
Archivist	4.4M	Django-workon	3.6M	PSPlot	2.7M	Phaser-ce	4.9M
Quanteda	3.7M	D1_python	3.5M	MIToS	2.5M	Phaser	4.7M
CollessLike	3.3M	Selenium	3.3M	PDSampler	2.2M	Ag-grid	3.6M
Readtext	3.1M	Senaite.core	3.3M	GeoIP	2.1M	Libsodium.js	3.5M
TOTAL R	942M	TOTAL Python	824M	TOTAL JULIA	264M	TOTAL CDNJS	1,323M

Outdegrees represent a one-way dependency relationship

A Sub-graph of the R Network: many packages depend on ggplot

OSS Packages with the Highest Dependencies

Selected network features and costs of top packages with the highest out-degree
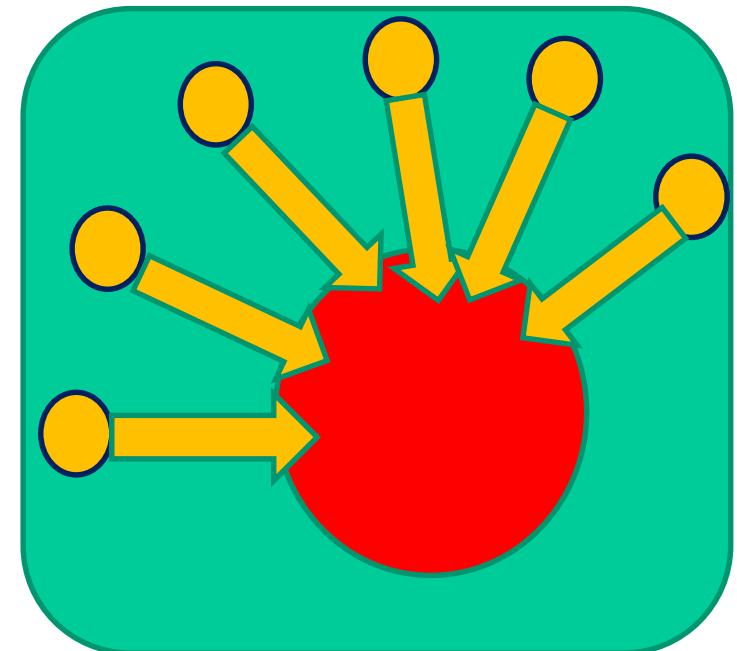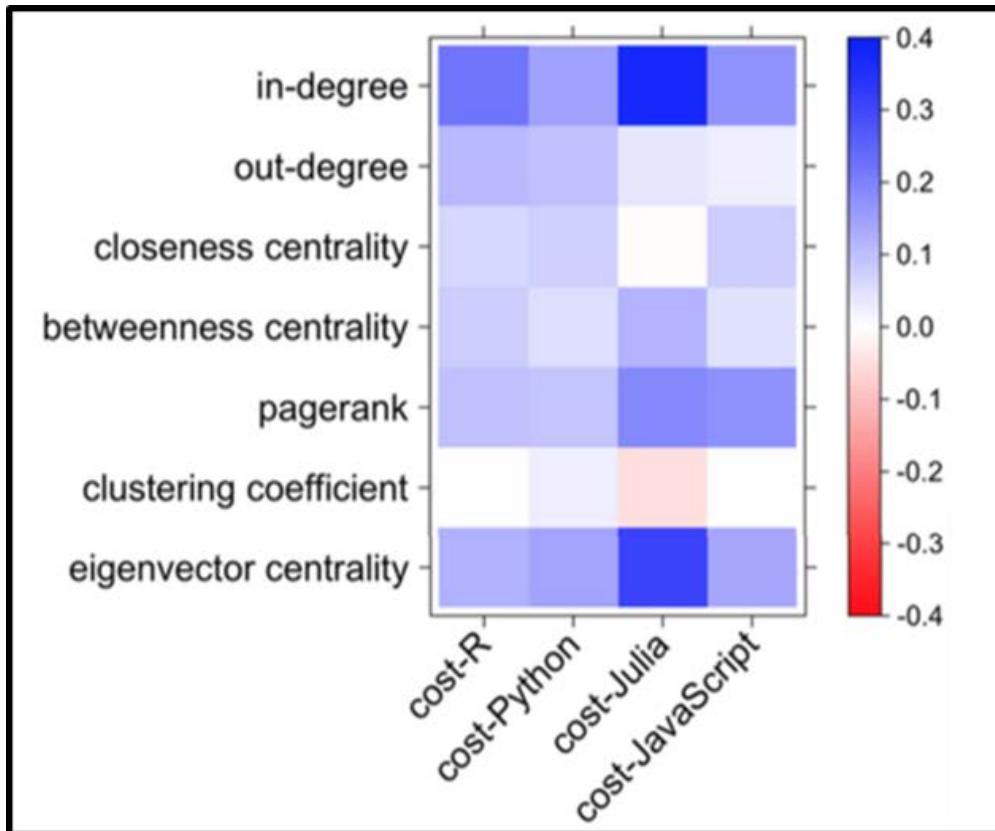
	Package	Outdegree	Indegree	Closeness	Between-ness	Eigen-centrality	Cost ($)
R	ggplot2	925	7	0.73	19.5K	0.07	1.08M
	Rcpp	838	0	0.50	0	0	960K
	dplyr	626	10	0.76	6.5K	0.06	875K
Python	requests	735	0	0.92	0	0	643K
	setuptools	182	0	0.71	0	0	555K
	scipy	131	0	0.96	0	0	2.58M
	Django	103	0	0.82	0	0	2.04M
Julia	Compat	596	0	0.60	0	0	235K
	Distributions	147	7	0.76	1.7K	0.07	530K
	StatsBase	136	4	0.56	856	0.02	306K
CDNJS	mocha	468	0	0.62	0	0	694K
	gulp	438	2	0.93	801	0.02	236K
	chai	258	0	0.86	0	0	633K

Top 5 packages with the highest total cost (in USD)

CRAN (R)	Cost	PyPi (Python)	Cost	Julia (Julia)	Cost	CDNJS (JavaScript)	Cost
Package	Cost	Package	Cost	Package	Cost	Package	Cost
googleAnalyticsR	4.4M	Nupic	3.9M	GeoStatsImages	4.2M	Webkit.js	7.6M
Archivist	4.4M	Django-workon	3.6M	PSPlot	2.7M	Phaser-ce	4.9M
Quanteda	3.7M	D1_python	3.5M	MIToS	2.5M	Phaser	4.7M
CollessLike	3.3M	Selenium	3.3M	PDSampler	2.2M	Ag-grid	3.6M
Readtext	3.1M	Senaite.core	3.3M	GeoIP	2.1M	Libsodium.js	3.5M
TOTAL R	942M	TOTAL Python	824M	TOTAL JULIA	264M	TOTAL CDNJS	1,323M

Relationship between Cost and Dependencies

Questions?
crobbins@nsf.gov

Grateful Acknowledgement to
Team Open Source Surfers (TOSS)
Virginia Tech Social and Decision Analytics Lab

Hannah Brinkley, Daniel Chen, José Bayoán Santiago Calderón, Eirik Iversen, Keren Chen, and Gizem Korkmaz