Modeling the Impact of R & Python Packages: Dependency and Contributor Networks

Gizem Korkmaz[1] **Claire Kelling**[2] **Carol Robbins**[3] **Sallie Keller**[1]

[1] Biocomplexity Institute, University of Virginia
[2] Penn State University
[3] National Science Foundation

2019 Women in Data Science Conference
Charlottesville, March 29, 2018

Introduction

- **Open Source Software (OSS)** is 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, 1998].
    - within and outside of the private sector
    - universities (e.g., Stanford, MIT, UC-Berkeley), business (e.g., Microsoft, Google), government research institutions (e.g., Sandia National Lab), nonprofits, and individuals
- Examples include Linux, Apache, Python, and R.
- Public funding for OSS is not fully accounted!

Introduction

Ways to Measure the Scope and Impact of Open Source Software

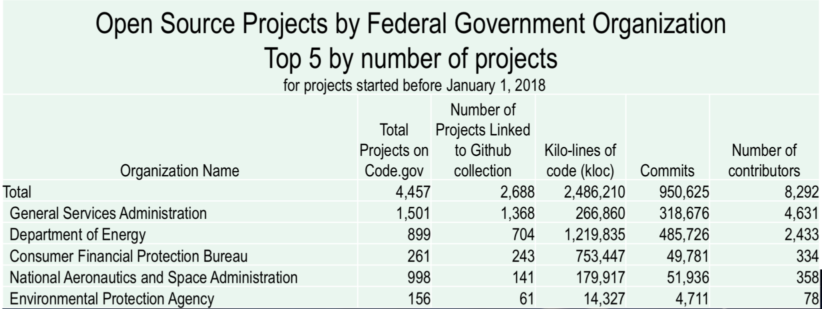
- Cost of Software Package Creation
  - Identify number of people involved each package's development;
  - Estimate time spent on software development;
  - Use average compensation for computer programmers;
  - Estimate intermediate inputs based on BEA (Bureau Economic Analysis) and OECD (Organisation for Economic Co-operation and Development) methodologies.

*The National Bureau of Economic Research (NBER) Conference: Big Data for 21st Century Economics. Bethesda, March 2019.

*International Monetary Fund (IMF) Statistical Forum on Measuring Economic Welfare in the Digital Age: What and How? Washington D.C., Nov. 2018.

*The International Association for Research in Income and Wealth (IARIW): The Digital Economy: Conceptual and Measurement Issues. Copenhagen, Aug. 2018.

OSS in Federal Government



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

Sharing America's Code

Unlock the tremendous potential of the Federal Government's software.

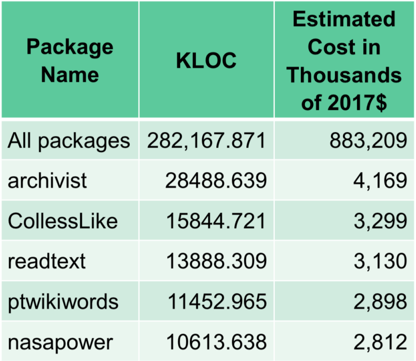
Search thousands of Federal Government projects

Ready. Set. Code!

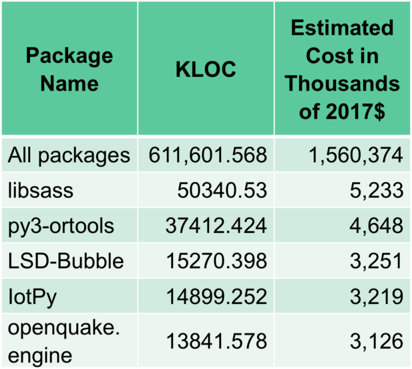
Whether you are a beginner or an experienced coder, join th
open source community. Help improve America's Code.

Resource Cost Based on GitHub Data

- Methods used in software engineering
- Based on lines of code, effort, and man-hours



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

- Refinement needed!

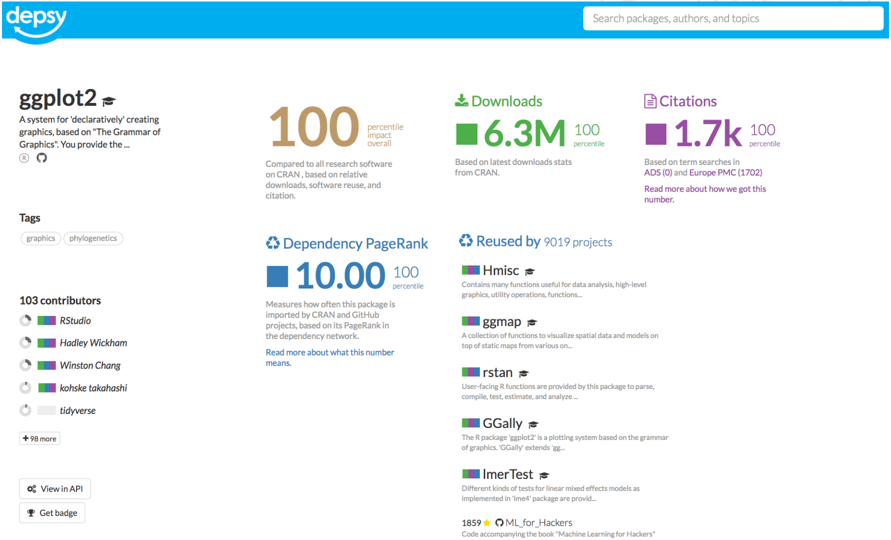
In this paper...

- Focus on **relative package impact (value)** of R and Python packages using methods from bibliometrics and patents
  - Counts
  - Downloads
  - Citations
  - Reuse in other packages (dependencies)
- Goal is to identify the factors that affect the impact, measured by **downloads and citations**
- Collect publicly available information on R and Python packages
- Generate **dependency and contributor networks** of OSS "ecosystem"
- Develop statistical models to estimate the impact using
  - structural properties of the dependency and contributor networks.
  - author and package attributes

*Preliminary version appeared in *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM)*. pp. 511-514. IEEE.

Data Source

Depsy.org: a website that compiles R and Python packages to quantify coding *impact* in the scientific community.



depsy

Search packages, authors, and topics

ggplot2

A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the ...

Tags

graphics phylogenetics

103 contributors

- RStudio
- Hadley Wickham
- Winston Chang
- kohske takahashi
- tidyverse

+ 98 more

View in API

Get badge

100 percentile impact overall

Compared to all research software on CRAN , based on relative downloads, software reuse, and citation.

Downloads

6.3M 100 percentile

Based on latest downloads stats from CRAN.

Citations

1.7k 100 percentile

Based on term searches in ADS (0) and Europe PMC (1702)

Read more about how we got this number.

Dependency PageRank

10.00 100 percentile

Measures how often this package is imported by CRAN and GitHub projects, based on its PageRank in the dependency network.

Read more about what this number means.

Reused by 9019 projects

Hmisc
Contains many functions useful for data analysis, high-level graphics, utility operations, functions...

ggmap
A collection of functions to visualize spatial data and models on top of static maps from various on...

rstan
User-facing R functions are provided by this package to parse, compile, test, estimate, and analyze ...

GGally
The R package 'ggplot2' is a plotting system based on the grammar of graphics. 'GGally' extends 'gg...

lmerTest
Different kinds of tests for linear mixed effects models as implemented in 'lme4' package are provid...

1859 ★ ML_for_Hackers
Code accompanying the book "Machine Learning for Hackers"

Data Collection

- **R**
  - Gathered all of the R packages listed on CRAN: 10,926 packages
  - Scraped the characteristics from the JSON page affiliated with each R package from Depsy
  - Information about 9.8K packages and around 14K affiliated contributors
- **Python**
  - Gathered all of the Python packages listed on PyPI: 192,666 packages
  - 43K Python packages and around 48K affiliated contributors

Contributor	Number of	Contributor	Number of
Name	Python Packages	Name	R Packages
Plone Collective	495	Hadley Wickham	104
ube	419	Scott Chamberlain	87
Amalgam8 Team	352	rOpenSci	74
Contributors	288	Dirk Eddelbuettel	64
Marc Abramowitz	288	RStudio	61
Total: 48,255 Contributors		**Total: 13,883 Contributors**	

Analysis

- Impact measures: citations and downloads.

Distributions of Downloads and Citations for Python and R

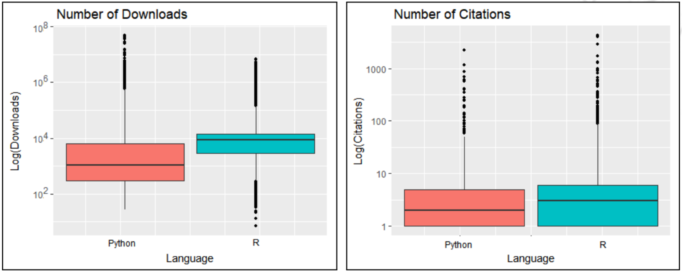


Figure: Histograms and boxplots of number of citations (left) and number of downloads (right) for Python and R packages. Both of these metrics are right-skewed.

Analysis

Top Downloaded Packages	Number of Downloads	Top Cited Packages	Number of Citations
Rcpp	6,683,565	vegan	4,275
ggplot2	6,255,500	lme4	4,023
stringr	5,366,703	nlme	2,916
plyr	5,345,308	ggplot2	1,702
digest	5,251,824	gplots	1,307

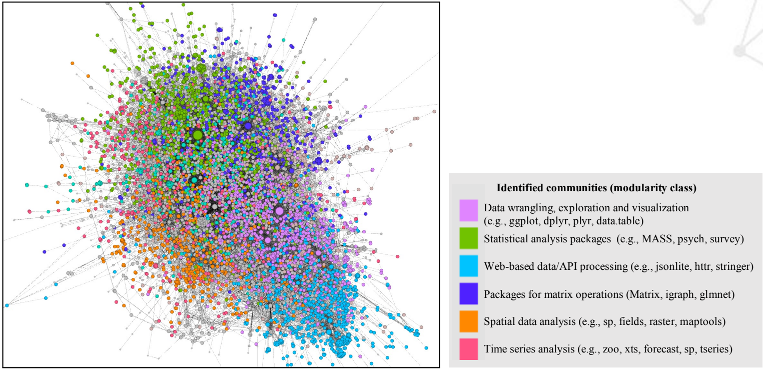
Table: **Top Downloaded and Cited Packages in R**

Dependency Network

- $i \rightarrow j$ indicates that the package j requires i to function

Table: **Structural Features of the Dependency Networks**

	Nodes	Edges	Avg. Degree	Diameter	Avg. Clustering Coefficient	Avg. Path Length	Connected Components
R	7,389	20,235	2.74	9	0.07	2.4	56
Python	168,921	545,186	3.23	22	0.07	7.8	174



Identified communities (modularity class)

🟪 Data wrangling, exploration and visualization (e.g., ggplot, dplyr, plyr, data.table)

🟩 Statistical analysis packages (e.g., MASS, psych, survey)

🟦 Web-based data/API processing (e.g., jsonlite, httr, stringer)

🟦 Packages for matrix operations (Matrix, igraph, glmnet)

🟧 Spatial data analysis (e.g., sp, fields, raster, maptools)

🟥 Time series analysis (e.g., zoo, xts, forecast, sp, tseries)

Dependency Network of R Packages

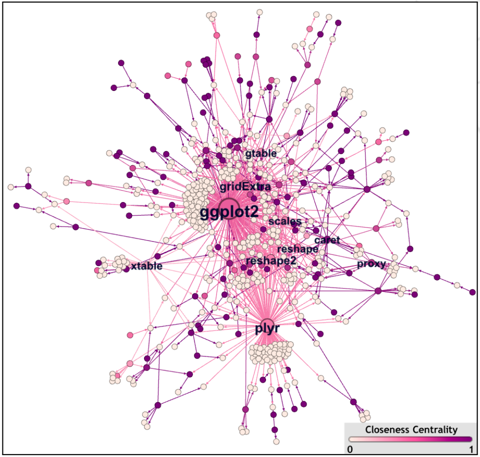


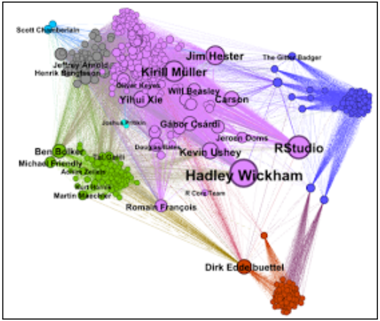
Figure: A subgraph of the dependency network of R packages on Depsy that includes 572 nodes (7.74%). The size of the node indicates the outdegree (the top nodes are labeled), and the color represents the closeness centrality.

Contributor Network

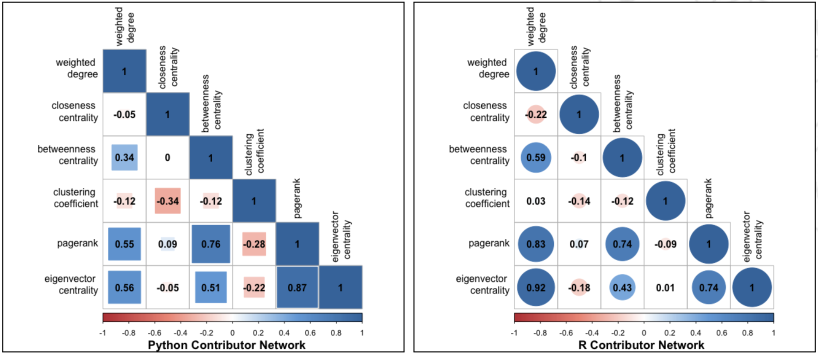
- An undirected edge between i and j indicates that user i and j contribute to the same R package (weight measures frequency).

Table: **Structural Features of the Contributor Networks**

	Nodes	Edges	Avg. Weighted Degree	Diameter	Avg. Clustering Coefficient	Avg. Path Length	Connected Components
R	12,340	90,030	15.4	19	0.84	4.4	1,651
Python	43,376	1,079,426	72.8	11	0.87	3.39	1,833



Centrality Measures and Correlations



Python Contributor Network

	weighted degree	closeness centrality	betweenness centrality	clustering coefficient	pagerank	eigenvector centrality
weighted degree	1					
closeness centrality	-0.05	1				
betweenness centrality	0.34	0	1			
clustering coefficient	-0.12	-0.34	-0.12	1		
pagerank	0.55	0.09	0.76	-0.28	1	
eigenvector centrality	0.56	-0.05	0.51	-0.22	0.87	1

R Contributor Network

	weighted degree	closeness centrality	betweenness centrality	clustering coefficient	pagerank	eigenvector centrality
weighted degree	1					
closeness centrality	-0.22	1				
betweenness centrality	0.59	-0.1	1			
clustering coefficient	0.03	-0.14	-0.12	1		
pagerank	0.83	0.07	0.74	-0.09	1	
eigenvector centrality	0.92	-0.18	0.43	0.01	0.74	1

Figure: Pearson correlation coefficients between the centrality measures of the Python and R contributor networks

Modeling Framework

- Two Quasi-Poisson models using downloads and citations as dependent the variables, y. We let $E(y) = \mu$ and $Var(y) = \theta\mu$. We assume that $y_i \sim Poisson(\mu_i, \theta)$ and let the mean μ_i for the i^{th} observation vary as a function of the p covariates as follows: $\mu_i = e^{\beta_0+\beta_1 x_{1,i}+\ldots+\beta_p x_{p,i}}$.
- p covariates: number of authors, number of commits, tag indicator, number of stars are used as package features, and network measures[1]
- For each package, we calculate weighted average of the degrees and centralities of the authors based on their percent contributions to the package.
- Three models are developed for each language.

[1] Some network characteristics were removed because of their high correlations with the other measures included in the models

Results - Full Model

		R	Python
	Variables	Estimate (sign.)	Estimate (sign.)
Package Feature	(Intercept)	9.97 (***)	9.32 (***)
	tag indicator	0.27 (***)	0.12 (***)
	number of stars	0.02 (.)	0.07 (***)
	number of commits	0.03 (*)	0.08 (***)
	number of authors	0.02 ( )	0.05 (**)
Dependency Network	indegree	-0.13 (**)	0.01( )
	outdegree	0.15 (***)	0.06(***)
	closeness centrality	0.39 (***)	0.35 (***)
	betweenness centrality	0.06 (***)	0.02 (**)
	eigencentrality	-0.06 ( )	-0.06 (.)
	pagerank	0.12 (***)	0.09 (***)
	clustering	-0.03 ( )	-0.20 (***)
Contributor Network	weighted degree	0.21 (***)	0.14 (***)
	closeness	-0.15 (***)	-0.11 (***)
	clustering	0.14 (***)	-0.16 (***)

Standardized coefficient estimates: (***) 0.001 (**) 0.01 (*) 0.05 (.) 0.1

Concluding Remarks

Summary of Findings:

- The more derivative a package is, the less likely it is to have a high impact
- Network centrality measures are significant
- Introducing network features improves model performance

Challenges:

- Citations of sofware packages
- Number of downloads is imperfect
- Data availability
- Representativeness

Next steps:

- Expand set of languages
- Incorporate cost of development

Thank you!

- Questions/Feedback?
- Contact:

Gizem Korkmaz (gkorkmaz@virginia.edu)
Research Associate Professor
Social and Decision Analytics Division (SDAD)
Biocomplexity Institute – University of Virginia

USDA United States Department of Agriculture

NSF

*This material is based on work supported by U.S. Department of Agriculture (58-3AEU-7-0074) and the National Science Foundation under IGERT Grant DGE-1144860, Big Data Social Science.