SDAL SOCIAL & DECISION ANALYTICS LABORATORY

DATA SCIENCE FOR THE PUBLIC GOOD

EXPLORING THE SCOPE AND IMPACT OF OPEN SOURCE SOFTWARE

Eirik Iversen (VT), Ben Swartz (VT), Claire Kelling (PSU), Sayali Phadke (PSU)
with Gizem Korkmaz and Stephanie Shipp (SDAL)
Sponsor: Carol Robbins, The National Center for Science & Engineering Statistics at NSF

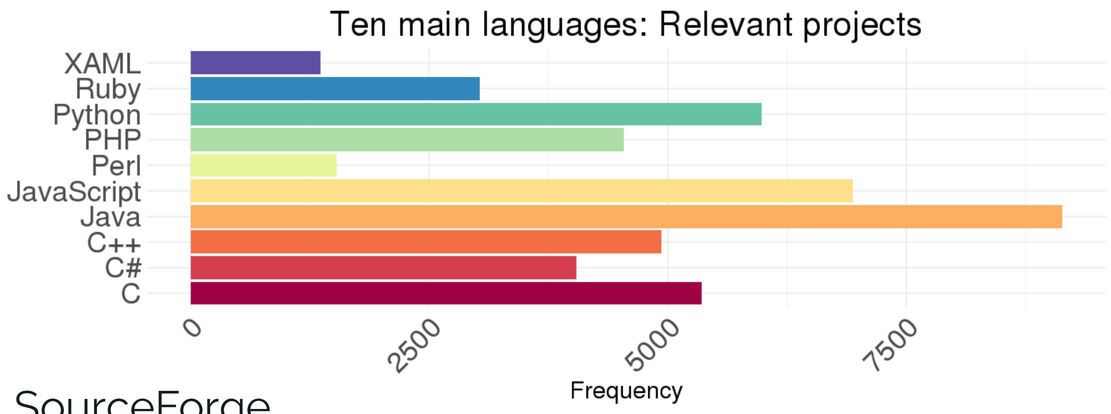
Project Description

The project aims to test the feasibility and to develop methods to measure the scope and impact (e.g., cost and benefit) of Open Source Software (OSS). Our goal is to understand how much OSS is in use (stock) and how much is created (flow). We use two data sources, OpenHub and SourceForge, to characterize the development and the categories (users) of OSS, respectively.

OpenHub

OpenHub is an online community and public directory of free and open source software. Source repositories include information on:

- Projects (~676,523): development info & activity, contributors, users, commits
- Organizations (698): portfolio projects (2850), contributors, organization info.
- People (~291,782): contributions, user information, project usage



We used their API (Application Programing Interface) to collect information on:

- a subset of the most relevant projects, determined by OpenHub (8.2% of all projects)
- a randomly selected subset (6.8% of all projects)

SourceForge

SourceForge is an "open source community resource" for development and distribution. It includes information on:

- Projects (~450,000) registered over the span of 18 years (1999-2017)
- Contributors (~350,000)

We used their API to collect the project names. Detailed information on each project was collected continuously over 17 days using web-scraping.

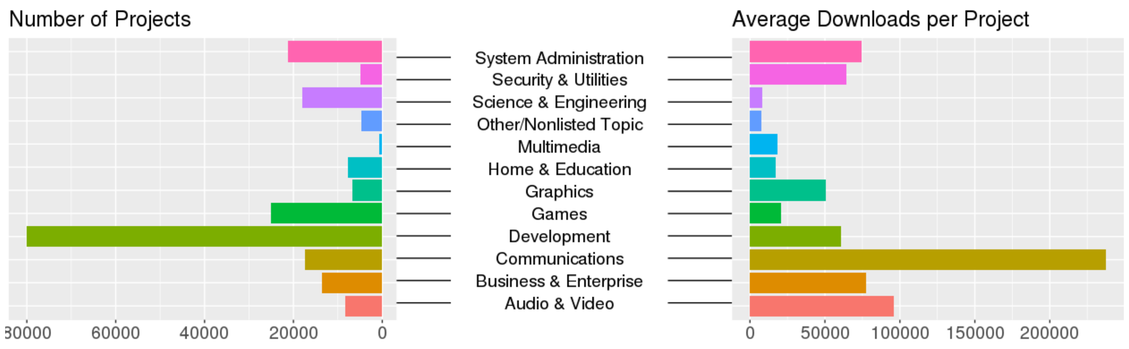
Top Downloaded Softwares	Category	Number of Downloads
Microsoft's TrueType core fonts	Development	2,079,351,723
Notepad++ Plugin Manager	-	1,481,067,341
VLC Media Player	Communications	898,907,923
Other Selected Softwares	**Category**	**Number of Downloads**
7-Zip (9th)	System Administration	412,199,932
Apache OpenOffice (11th)	Development	238,646,302
PDF Creator (16th)	Business & Enterprise	113,284,095
Weka -machine learning- (154th)	Science & Engineering	8,797,668

Categories and Subcategories

Projects in SourceForge are structured with three levels of categories. Categories and respective top subcategories are given in the table on the right.

The figure below illustrates that the number of projects per category and the average download by category vary.

Category	Top Subcategories
Audio & Video	Sound/Audio (68%), Video (27%)
Business & Enterprise	Enterprise (27%), Financial (24%)
Communications	Chat (32%), Email (21%)
Development	WWW/HTTP (26%), Software Development (8%)
Games	Games/ Entertainment (23%), Role-Playing (13%)
Graphics	3D Rendering (19%), 3D Modeling (12%)
Home & Education	Education (39%), Computer Aided Instruction (15%)
Science & Engineering	Bio-Informatics (14%), Artificial Intelligence (12%)
Security & Utilities	Security (59%), Cryptography (35%)
System Administration	Networking (19%), Storage (15%)



Network of Categories

We use the category and subcategories of each project on SourceForge to generate the network of categories.

The full network, shown on the right, includes 346 nodes (i.e., categories), and 467 directed edges.

The full network above provides an impression of the overall scope, and the subnetwork of Science and Engineering is presented on the left.

Challenges and Next Steps

- Challenges include computational complexity and data quality issues.
- Future work involves collecting a larger dataset from OpenHub and developing methods to measure the impact of OSS.