EVALUATING INNOVATION OF OPEN SOURCE SOFTWARE

Romcholo Macatula, John Higgins, and Alex Gagliano
In Collaboration with Dr. Gizem Korkmaz and Dr. Stephanie Shipp
The Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech
Computational Modeling & Data Analytics, Virginia Tech

Research Question

The National Science Foundation (NSF) is interested in measuring innovation using non-survey sources of data that will allow them the opportunity to explore innovation in new areas, such as the contribution due to Open Source Software (OSS). This research aims to develop a series of metrics to quantify this contribution. Methods include regression, Markov chains, and several unsupervised methods to study the spread of and current trends in OSS. Data will be scraped from various sites that compile information about OSS. Overall, this research will be a first step for NSF's adaptation into funding OSS.

Data

We hope to find a set of relatively easy to collect data that can track the level of innovation of certain OSS. We aim to scrape data on the following sources,

➢ ***Social Media***
- Only has data on larger OSS
- Look at very deep insights on user interactions
  - User involvement & sentiment analysis

➢ **Sourceforge.net**
- Data on a multitude of OSS
  - Weekly downloads, ratings
  - Category of software
  - Includes the most number of OSS

➢ **Blackducksoftware.com**
- Database on OSS & community projects
  - Deeper insights with smaller communities

Methods

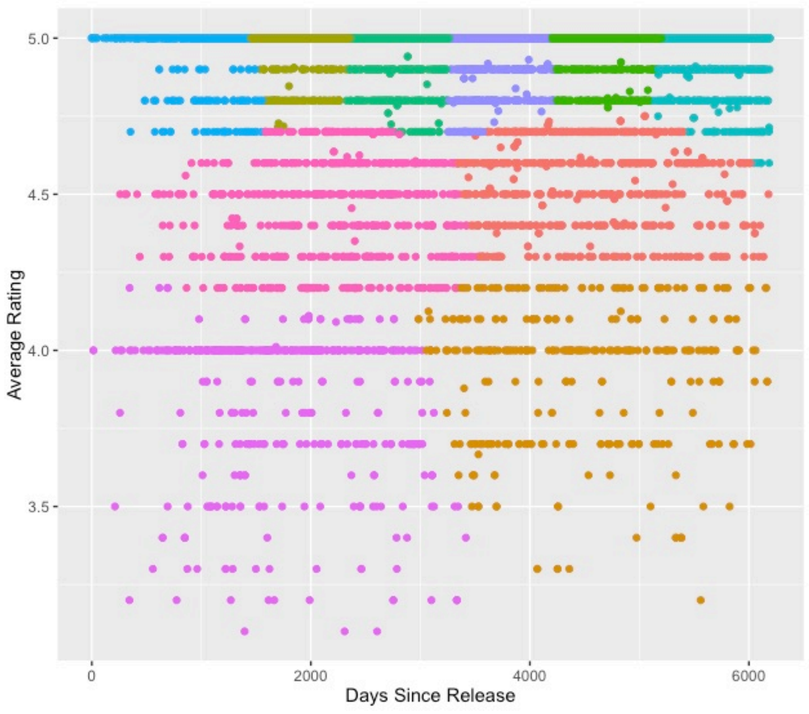
K-Means
- Discovering underlying but non-trivial clusters of data

Evidence Gap Map
- Graphically explore explanatory & response factors through a literature review
- Create matrix which visually graphs individual studies and papers

Analysis

K-Means Clustering – Discovering underlying but non-trivial clusters

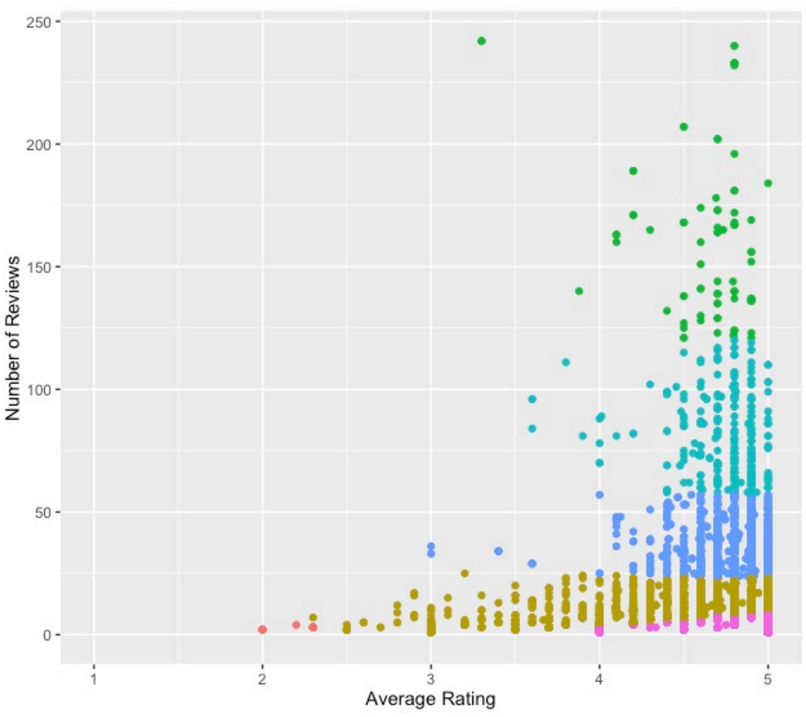
Average Rating vs. Days Since Release



Average Rating vs. Days Since Release (Left)
- Note that there is no correlation between the two factors
  - Common trend among factor combinations
- Performed K-Means with 10 clusters
- Top left cluster (Light Blue) groups OSS with relatively *new* software and *high ratings*
- Potentially indicates emerging, innovating software

Number of Reviews vs. Average Rating (Right)
- Appears to be a pattern in the data
  - Potentially consider data as censored
  - High reviews correlate with more reviews
- Bottom right cluster (Pink) groups OSS with *high ratings* and *low* number of reviews
- Potentially indicates a high performing OSS within a small niche
  - Could also indicate OSS that has the potential for innovation with more growth and development

Number of Reviews vs. Average Rating



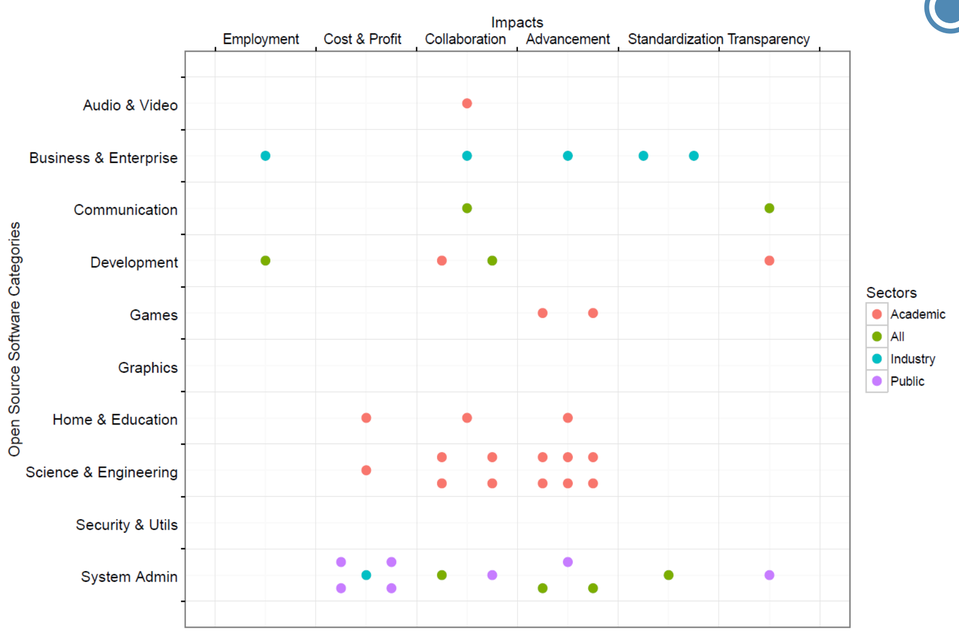
Evidence Gap Map – Graphically exploring the literature

Description of Evidence Gap Map
- Rows: Fields of OSS
  - Due to categorization in our data
- Columns: Major Impacts
  - Chosen through literature review
- Colors: Economic sectors of OSS
- Each dot represents a paper documenting the relationship

Major Feature: Academic Sector
- Predominant sector of documented OSS impacts
- Major impacts within Collaboration and Advancement
- Majority of OSS in Academia potentially in Science & Engineering, and Home & Education
- Reflects the basic notions around the scientific community

Major Feature: System Administration
- Lots of studies focused on this field
- Second largest portion of OSS in Industry potentially within System Administration
- Almost all literature on OSS in Public sector within System Administration
- Many studies document the cost benefits of this type of OSS

Evidence Gap Map



Overall Features
- Across sectors, the majority of studies on OSS look at the impact on Collaboration and Advancement
  - Perhaps this points to the two largest goals of OSS
- We note a lack of documentation within Employment, Standardization, and Transparency
- Some OSS do not lend themselves to academic studies, such as OSS in the field of Graphics, Games, and Security & Utilities
  - Perhaps there are undocumented impacts which require further research

Conclusions & Next Steps

Through our first sweep of exploratory methods, we begin to evaluate the current climate for OSS. We see hints of the interactions between OSS, sectors of the economy, and the characteristics within each sector in our Evidence Gap Map. Through clustering, we identify potential metrics in evaluating innovation in OSS.

The largest challenge moving forward is the identification and collection of data. It is difficult to find a database of OSS that contains all the information we are seeking. While sourceforge.net has data on a large swath of OSS, the data collected is fairly general. When looking at social media data, we have the opposite problem: very specific data on a few OSS.

In the future, we hope to find more connections between our factors such as software type, field of predominant presence, and user activity. Considered methods include Hierarchical Clustering, ANOVA, and other Association Analysis methodologies.

Finally, we wish to further identify important qualities of OSS that suggest innovation within their field. Not only that, we hope to quantify and trace qualities of innovation through specifically numerical data. Perhaps developed any metrics from this study can be used to supplement traditional means of measuring innovation.

Acknowledgement

We acknowledge Biocomplexity Institute of Virginia Tech, the Division of Computational Modeling and Data Analytics, and the Social & Decision Analytics Laboratory for this opportunity. This material is based upon work supported by NSF under Cooperative Agreement #1641251.