Mining GitHub
Inside the R Community

Abstract
Since its inception in 2008, GitHub has come to house over 1,000 R-language projects, including over 180 CRAN packages. In this article, we investigate the social structure of the R community on GitHub using the GitHub API, an open web programming interface to GitHub's repositories and metadata. This generic API includes search facilities which can enumerate repositories associated with specific programming languages. This interface will allow us to investigate the landscape of R development on GitHub and identify key projects with the greatest social impact to the community. Having identified projects of interest, we can use additional facilities of the API to harvest related development activity data, which we can then visualize to better understand the scope of collaboration within the associated repositories.

1. How does R compare to other languages on GitHub?

2. What R repositories on GitHub attract the most attention?

The top 10 most frequently forked and watched repositories from the Python, Java, Scala, and R communities were identified. The associated charts demonstrate that the R community is relatively small compared to these more general-purpose programming language communities.

3. Is the development effort well distributed in the R community? What level of collaboration is present on the most influential projects?

4. Do people pay more attention to CRAN packages?

5. How active are the projects on GitHub?

Based on the nominal line shown in blue, there is a collection of projects that are very active. Whereas, those along the x-axis have seen little or no activity since they were first created and added to GitHub.

6. How well are CRAN packages represented on GitHub?

Although limited in number, CRAN packages contribute a substantial portion of the repositories with high fork counts. The small multiples in this visualization demonstrate the distribution at different zoom levels, while the x-axis shows the respective R versions at the time of the project's creation.

Tools Used on the Project
(a) The project was developed using R version 3.0.1
(b) Data collection was performed using the RCurl and rjson packages
(c) Data manipulation was performed using plyr
(d) Visualizations were produced using ggplot2

Patrick Toohey
Department of Human-Computer Interaction, SUNY Oswego
ptooley@oswego.edu