DS 6040 Bayesian Machine Learning
Final Project

Open Source Software Network Analysis

Team

Derek Banks (dmb3ey), Camille Leonard (cvl7qu), Shilpa Narayan (smn7ba) and Nick Thompson (nat3fa).
We are also part of the Open Source Software Network Analysis Capstone Project.

Problem Statement and Background

Abstract: Open Source Software (OSS) is computer software with its source code shared with a license in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose. Examples include Linux operating system, Apache server software, and R statistical programming software. Despite its extensive use, reliable measures of the scope and impact of OSS are scarce. The creation and use of OSS highlight an aspect of technology diffusion and flow that is not captured in science and technology indicators. Supported by the National Science Foundation (NSF) and building on research conducted over the last couple of years, we aim to measure the production, impact, and diffusion of OSS in specific sectors, institutions and geographic areas using data scraped from multiple hosting platforms (e.g., GitHub, GitLab, SourceForge). We will generate and analyze networks of contributors (through collaborations between software developers) and networks of OSS projects (through reuses across projects and shared contributors), and will identify key/influential players in this ecosystem.

Data Sources

Python package information and relevant metadata will be downloaded from the Google Big Query API. Additional data can be sourced from the PyPI API. A manageable subset of the downloaded data will be used for this project.

Objective

For the capstone project, our group will be focusing on conducting network modeling of OSS licensed python packages. For this final project, we will be using Bayesian methods to predict linkages between nodes. Should time allow we will also explore community prediction using Bayesian methods.

Data Analysis Tools

Libraries from python programming languages such as pandas, numpy, scipy, statistics and others as required. We will program in Jupyter notebooks. We may use other tools or packages outside of python to gather data, clean or process as required and will continue to update this list.

Data Products

A constructed network with linkage and community predictions.