

# Densities of Contribution Networks for Open Source Python Packages

Rey Koki

October 15<sup>th</sup>, 2021

## Proposal

Successful open source software (OSS) projects often obtain more contributors as their popularity increases, but they vary in the amount of effort they put into promoting contributions from outside parties. Numpy, while recently criticized for its lack of diverse contributors, provides potential contributors a 'good first issue' tag. On the otherhand, JupyterNotebooks gives instructions for creating a developer environment but no simple guidelines on how to initially contribute to the project.

We want to investigate the density of the networks of individual python software projects and their popularity measured by the number of downloads. For each software project network, there will be two types of nodes, the project itself and the contributors which will all have a single edge that connects to the project node. The weight of each edge will be equal to the number of commits the user has contributed to the master branch. In terms of 'density' the larger the edge weight, the closer it is to the center project node.

The anticipated findings will be that the more popular the OSS python project, the more single commit contributors it will have. The correlation between popularity and 'core' contributors is more difficult to predict, the initial team members may be able to focus solely on the project as it gains funding or they could have moved on to other projects or moved to more of moderator positions, thus making less commits. If we find that some projects with similar popularity have very different results, it would be interesting to investigate the language used in their collaboration portion of their GitHub repo. In addition, correlations between application type could be studied, the networks for pytorch and keras might have different properties than the networks for scipy and numpy. The hope of this study will be to gain further insight on how to increase contributions, and hopefully could be expanded to increase commits from a more diverse group of contributors.

The GitHub API is publically accessible and able to provide the project contribution data needed to create the networks. The popularity of the python projects will be determined by either the number of forks of the repo or the number of user installs. Networkx is able to create the networks, rediculogram representations and produce the results needed for the analysis portion of the project.