# VANT 149 Research Project Proposal: Amount of Developers Influenced Size of OSS Projects

Kewei Li and Yuhang Zeng

Vantage College, University of British Columbia

## Abstract

Open Source Software (OSS) has recently being increasingly popular. OSS are usually being cooperate coding on OSS communities such as GitHub, SourceForge etc. GitHub has become one of the biggest OSS community, by September 2017, GitHub owned almost 24 million users and 67 million repositories (GitHub, 2017). The current study is devoted to explore the relationship between the number of developers and the size of OSS project on GitHub platform, since there exists no recent study focused on this relationship. In this study we will develop a program running on GitHub to uncover this relationship by analyzing ten percent of all newly built GitHub repositories between September 2016 and September 2017 will be analyzed. The verification of this relationship can potentially benefit the OSS communities and software companies by giving them a general understanding of relationship between developers and size of software project.

# Amount of Developers Influenced Size of OSS Projects

## Introduction

With the development of Internet and computer technology, the demand of application of programs are getting more and more complex. These complex applications also lead program designing problem more and more complex. With this developing difficulty, a new form of software came into being: Open-source software (OSS). OSS is a kind of software “under a special software license that allows users to use, change, and/or improve the software's source code, and to redistribute the software either before or after it has been modified” (Mayhew, 2015). People can contribute towards a same project, and everyone can do what they are good at. The appearance of OSS solved the increasingly developed demand, the complexity coding load may be hard to deal with individually but the workload can be simplified if a number of developers are involved. One of the most famous OSS community is GitHub. GitHub is an online platform which is widely used among all multi-person collaboration needed fields such as teamwork program develop. It stores the projects on its server so that multi people can work together because all the people who work with it can have access and edit. Other features such as access control, bug tracking, feature requests, task management, and wikis are offered for every project. By September 2017, GitHub owned almost 24 million users and 67 million repositories (GitHub, 2017), making it one of the largest source code database in the world. Most of previous research in the field of OSS, has a lot of limitations (will specify in Literature Review section) and no recent study have focused on the relationship between the number of contributor and the size of OSS project. Thus, in order to fill the gap of knowledge, in this study, we will explore the relationship between the size of repositories (refers to the overall size of source code) and the amount of developers involved by analyzing 10% of the newly built repositories between

September 2016 and September 2017. This study is devoted to indicate this relationship. The verification of this relationship can potentially benefit the OSS communities and software companies, it may give them a general understanding of relationship of developers and size of software project.

## Objective

Explore the relationship between the size of OSS project and numbers of developers.

## Literature Review

With the increasingly popular of Open Source Software (OSS), and its public access permit, it had drawn researchers' interest. A series of previous research had been done and give the current study a general idea. A 2016 study is focused on how outer developer's participation related to the prospect of the project (Krishnamurthy, Jacob, Radhakrishnan & Dogan, 2016). The authors did the following experiment: 2600 archives have been analyzed and 7 factors was controlled (number of users, code complexity, license restrictiveness, age of project, number of bugs, size of project and programming language). This 2016 study has limitations in the number of archives and also controlled too many factors. The current study was designed to take much more amount of sample archive size and will not control those factors, except the age of project.

Another previous study worth mention is a study done by 2006, focused on how motivations of contributors of OSS are being related (Roberts, Hann & Slaughter, 2006). This study analyzed projects under Apache license from 1999 to 2002, by using an original theoretical model with Structural Equation Model (SEM) (Roberts, Hann & Slaughter, 2006). This 2006 study was using data from 1999-2002, which is relatively an old data set which made it a limitation of the study. To avoid the same limitation, in this study, we decided to use new data between 2016 and 2017.

## Research Plan

According to the official report given by GitHub, 25 million activate public repositories from September 2016 to September 2017 (GitHub, 2017). Activate had been defined as at least one public activity have been done during this time (GitHub, 2017). Given these official data, we prepare to analyze 10% of overall 25 million repositories which is 2.5 million repositories.

### I. Program Develop

The following module will be developed in order to finish the research program.

GITHUB ACCESS MODULE

This module is designed to provide this research full access of GitHub API access. A secure token (provided by GitHub) is employed, and this personal secure token can provide a maximum 5000 requests per hour, while normal request is limited in 60 requests per hour. All information gaining process, or “crawl” called in the following part of the proposal, will be process by next module.

USER-GRABBING MODULE

This module consumes a group of seed users from GitHub, and crawl the followers of those seed users, and keep going to crawl the followers of the followers of seed users. A natural recursion should be employed here. And all the followers will be stored in a multi-tree structure. Following operation on followers in the multi-tree will be handled in the next module

REPOSITORY-STORAGE MODULE

This module will consume the followers from the multi-tree which USER-GRABBING MODULE generated. And will crawl all the repositories built by current user, with the same time, the creation date will be recorded, if the date of creation was after September 2016 and before September 2017, and also satisfy that the last update date is differ from the create date (i.e. the repository has been changed after it been created.), this repository will be downloaded and will be given to the next module to handle.

REPOSITORY-ANALYSIS MODULE

This module will consume the repositories downloaded by the last module, and by using command “git shortlog --summary --numbered” to generate the developers of the current repository. Then the number of developers will be storage and the size of repository will also be stored. All gained [number of developer] - [repository size] will be stored in database.

## II. Data Analysis

The gained data will be analyzed by the following method.

The data gained from the previously designed will employed to find a pattern between the size of OSS project and numbers of developers. It has been hypothesized that with the increase of size of OSS project, the numbers of developers will increase. However, the trend of increase is unclear. Hence, we will try to match the data into six different trendline functions, including 1. Linear, 2. Logarithmic, 3. Polynomial, 4. Power, 5. Exponential and 6. Moving average (Microsoft, 2018).

## III. Timeline

As a summary, our research will be done by five stages (time detailed):

Stage 1: Setup and fully test the program (design has been given above) which will be employed in this study (1st May to 19th May). (Done)

Stage 2: Test the runtime and estimate the required compute resource and ask UBC CS department for additional compute resource support (19th May to 21st May). (Done)

Stage 3: Run program (22nd May to 12nd June).

Stage 4: Analyze the data (13rd June to 23rd June).

Stage 5: Prepare for presentation (23rd June to 30th June).

# References

GitHub. (2017). *GitHub Octoverse 2017 | Highlights from the last twelve months*. Retrieved from https://octoverse.github.com/

Krishnamurthy, R., Jacob, V., Radhakrishnan, S., & Dogan, K. (2016). Peripheral developer participation in open source projects: an empirical analysis. *ACM Transactions on Management Information Systems (TMIS), 6*(4), 14.

Mayhew, S. (2015). open source software. In A Dictionary of Geography. Retrieved from http://www.oxfordreference.com/view/10.1093/acref/9780199680856.001.0001/acref-9780199680856-e-4035

Microsoft. (2018). *Choosing the best trendline for your data - Office Support*. Retrieved from https://support.office.com/en-us/article/choosing-the-best-trendline-for-your-data-1bb3c9e7-0280-45b5-9ab0-d0c93161daa8

Roberts, J. A., Hann, I. H., & Slaughter, S. A. (2006). Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects. *Management science, 52*(7), 984-999.