DREU Proposal

Rebekah Manweiler

May 18 - Aug 3, 2018

1 Introduction

1.1 Project Details

This is a working document for my Distributed Research Experience for Undergraduates funded by the CRA-W. This summer I am working with Professor Nicole Beckage and Professor Joe Austerweil at the University of Wisconsin - Madison. They are working on the DARPA Cyber Grand Challenge where the challenge is to model different aspects of the Software Social Network (SSN) GitHub. They are currently working on the baseline challenge which is designed to focus on "simulating social structure and temporal dynamics of key processes, as well as individual, community and population level behavior on GitHub." They have been given GitHub user and repository data and five high level research questions. These research questions are further specified with 31 different questions and related metrics.

- RQ1: How do users **engage** in the development of technologies and how does their engagement evolve?
- RQ2: How do users **contribute** to the evolution of technologies; how much, how quickly, and how evenly?
- RQ3: How do users' levels of **trust and reputation** impact the rate of innovation growth?
- RQ4: How do contributions to one technology **influence** the development of other technologies?
- RQ5: How quickly do technologies become **popular** and how does their popularity evolve over time?

My work this summer will be centered around this project. I will be using the MySQL GHTorent database to model users and their actions on GitHub. From the list of metrics for the main project, I categorized each question by determining if it focused on a user, repository, or community, and if it focused on a general model or comparative model. The questions and related metrics I selected for my project were questions I categorized as general user model questions.

- (27 from RQ1) How soon do users engage with GitHub over time after joining?
  Metric - Diffusion delay of user actions (excluding Fork and Watch events) since the creation of the user account.
- (30 from RQ1) Do users who initially engage with repositories continue to contribute over time?
  Metric - Probability that users will continue to contribute to the repository given how many previous interactions with that repository the user had.
- (31 from RQ1) How much activity is required before the users become committed project contributors?
  Metric - The ratio of the probability that a user's next action will be an Issues event to the probability that it will be a Push event as a function of how many previous interactions the user has had with a repository.
- (5 from RQ2) What are typical patterns of activity observed for developers on GitHub?
  Metric - Distribution of total events by week day/hour.
- (17 from RQ2) Do users contribute across many repositories?
  Metric - Number of unique repositories that users contribute to.

- (24 from RQ2) What are the basic characteristics of the developers' population?
  Metric - Distribution over user activity for all users. Top K most active users (total number of actions per user).

My goal with these questions is to understand how people are using GitHub and how their usage changes over time as well as the different kinds of GitHub users and the differences in their behavior. We already have an intuition for the different kinds of GitHub users like students, professors, small businesses, and companies, and we can imagine how they might use them differently as a starting point. For instance, students may use their GitHub accounts more sporadically and more exclusively for school projects with other students while developers at a company may use their accounts more regularly in conjunction with a software development framework and for multiple ongoing projects. But, I want to know if we can mathematically detect and categorize those differences to generate a basis for all the current and potential users of GitHub and from there be able to more accurately model and predict user behavior.

2 Cognition

3 Models

3.1 Bayesian Models

3.2 Poisson Processes

3.3 Hawkes Processes

4 GitHub

GitHub has been described as an Open Source Software (OSS) community and Software Social Network (SSN) in recent publications. GitHub is a network of public 'repositories' or projects which can contain anything from code to executables or documents to images. GitHub is generally used as a platform to share and manage coding projects including software for games or applications, libraries for coding languages, manuals and specifications for software or hardware, and much more. GitHub was built off of and can be used with the document source control system Git which allows users to manage and edit different versions of one document. GitHub stores a copy on-line of any folder or file system that is being monitored by Git: these are a user's repositories. Repositories and local folders are updated by using commands like 'push' or 'pull'. GitHub is largely used by groups or teams of users that are all working on the same project. Users will all have a local copy of the repository which they can work on and edit, and then when they want to share their changes with the group they can 'push' their changes to the repository and tell other users to 'pull' their changes from the repository. Users can also add comments to their changes to give notes or explanations for their changes. In this way, groups can easily edit and share a project.
Additions to this environment has allowed it to take on an aspect of social media. Users on GitHub can follow the activity of other users, watch the activity on a repository from many users and receive notifications for activity, star certain projects to bookmark them for later or receive more limited notifications on repository activity, fork a users project to create their own local copy, create pull requests for repositories they do not own to ask for their changes to be included in a project, and several more. These additions create an environment for users to interact with each other while developing and managing software, of which cognitive scientists and the like are interested in understanding.

4.1 Prior Research

In the recent years, research into the behavior of users on social networking sites has sparked a similar interest in social coding platforms like GitHub, Stack Overflow, HTML5 Rocks, and many more. These websites allow software developers to collaborate on projects and share code easily, creating a community rich in resources and ideas. Many social scientists wish to understand how these communities are created, how they thrive, how users interact with one another, and how they influence each other.

4.2 Project Approach