USER INTERACTION BEHAVIOR IN GITHUB ORGANIZATIONS: A CASE STUDY ON REDDIT AND 10GEN

Parth M. Shah[1], Biswadeep Khan[2], Prajata Roy[3], Amit A. Nanavati[1] and Hridoy Sankar Dutta[4]

[1] Ahmedabad University, India [2] Hong Kong University of Science and Technology, HK
[3] NIT Durgapur, India [4] University of Cambridge, UK
parth.s5@ahduni.edu.in, bkhan@connect.ust.hk, prajata111@gmail.com, amit.nanavati@ahduni.edu.in, hsd30@cam.ac.uk

Abstract

GitHub organizations help developers to collaborate across repositories in an organization. Past studies on GitHub focused on user or repository-centric analysis with objectives such as evaluating contributions [1], code quality [2], identifying code duplicates [3] etc. However, only a limited number of them study GitHub organizations. Our final goal is to recommend users to repositories in the context of a GitHub organization, and so we are studying user-repository interactions in such a setting. In this paper, we discuss the insights obtained by analyzing two GitHub organizations' network structures.

Our paper has the following contributions:
1. We plan to release novel graph datasets[1] of interactions in GitHub organizations. This, to our knowledge, is the first dataset of such kind.
2. The analyses of the network properties of two types of user-repository interactions[2]: *stars* and *watches*, and their comparison in GitHub organizations.

Please consider this submission either as a oral presentation or a poster presentation.

Data collection & Graph construction

We used the GitHub API[3] to collect the information on GitHub organizations. We identified that GitHub organizations are maintained with incremental identifiers starting from 1. We collected data for the first 50k organizations, however we observed that most of these organizations have very less number of users and repositories. To speed up the data collection process, we parallelize the use of the API keys using Python Joblib and multiprocessing library.

Based on the extracted data of GitHub networks from different organizations, each network can be represented as a bipartite network of users and repositories based on the actions performed by the users on these repositories: *stars* (B_s) and *watches* (B_w) where an edge specifies the user has starred (watched) the respective repository. Further, we generated the two unipartite user-user projection graphs: *stars* (G_s) and *watches* (G_w) from B_s and B_w respectively, where an edge specifies that they starred (watched) at least one repository in common. For this work, we study two organizations: `Reddit` ($|V|_{B_s}$=34913, $|E|_{B_s}$=43971, $|V|_{B_w}$=45053, $|E|_{B_w}$=44553) and `10gen` ($|V|_{B_s}$=3789, $|E|_{B_s}$=3950, $|V|_{B_w}$=4331, $|E|_{B_w}$=4283). We used Gephi and Python NetworkX library for our analysis.

	$Reddit_{B_s}$	$Reddit_{B_w}$	$10gen_{B_s}$	$10gen_{B_w}$
APL	6.811	13.79	7.544	8.199
ACC	0.6583	0.9223	0.823	0.911
Diameter	14	40	17	20
Modularity	0.47	0.94	0.667	0.768

Table 1: Network properties of B_s and B_w. APL refers to Average Path Length and ACC refers to Average Clustering Coefficient.

Preliminary observations & Future work

The analysis on the network structures of two organizations provides the following observations:

1. **"Star"-ing is a more passive user action compared to "Watch"-ing.** (see Table 1): We observed that APL, ACC and Diameter for B_w is greater than B_s across both the organisation.
2. The modularities (last row in Table 1) of B_s and B_w (treated as unipartite graphs) shows that **`Reddit` has a more pronounced community structure than `10gen`**(users are less scattered across different sets of repositories) .

We plan to extend this analyses to G_s and G_w in order to develop a recommendation system for GitHub projects. We also plan to release an extended version of our graph datasets as a bi-product of this study.

References

[1] Jason Tsay et al. Let's talk about it: evaluating contributions through discussion in github. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*, pages 144–154, 2014.

[2] Baishakhi Ray et al. A large scale study of programming languages and code quality in github. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*, pages 155–165, 2014.

[3] Cristina V Lopes et al. Déjàvu: a map of code duplicates on github. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA):1–28, 2017.

[1] https://tinyurl.com/github-organizations

[2] Stars (Watches): when a user likes your repository or they want to show some appreciation (wants to be notified of all the activities in a repository), they star (watch) it.

[3] https://docs.github.com/en/rest