A Partial Reproduction of Tsay, Dabbish, and Herbsleb's (2014) Influence on Social and Technical Factors for Evaluating Contribution in GitHub

Rahul N. Iyer[1], S. Alex Yun[2,*], Meiyappan Nagappan[3], Jesse Hoey[4]

[1] r3iyer@uwaterloo.ca, githubID: rahul-iyer
[2] alex.yun@uwaterloo.ca, githubID: salexyun, corresponding author
[3] mei.nagappan@uwaterloo.ca, githubID: meido
[4] jhoey@cs.uwaterloo.ca, githubID: jessehoey

Link: https://www.dropbox.com/sh/fn24okch8aeu763/AAClp1S4bY4T0Ww-vh4zkHora?dl=0

We reproduced Tsay, Dabbish, and Herbsleb's [1] work that examined the influence of social factors and technical factors for evaluating GitHub contributions. Social factors are measures that are related to one's social connections, which include social distance, prior interaction, and followers. As the name suggests, technical factors include test file present, total churn, files changed, number of comments, main team member, team size, stars, and project age.

While it is evident that one's technical merit plays an important role on the likelihood of contribution acceptance, the role of social factors on open source software project contributions were relatively unknown until Tsay et al.'s work.

In theory, identifying these factors could delineate how important decisions are made in both online and offline collaborative settings. This also has practical implications where developers can think of ways to maximize their contributions by reflecting on how their work is going to be evaluated by fellow developers. We were motivated to reproduce Tsay et al.'s work on a more recently mined dataset to determine if their results could be generalized.

Tsay et al. created a dataset of pull requests, users, and repositories, using the GitHub API and the GitHub Archive dataset. After two phases of sample filtering, they were left with 12,482 projects and 95,270 developers.

We took 11,000 projects used by Tsay et al. and supplemented with the *RepoReaper* dataset curated by Muniah, Kroch, Cabrey, and Nagappan [2], which included more than one million projects. We further filtered the *RepoReaper* dataset down to 15,000 projects. After the additional filtering, our final dataset included 1,860 projects and 16,935 developers.

We reproduced the effects of social and technical factors on pull request acceptance. All factors except for follower count had significant effects comparable to that of Tsay et al.'s work. Social factors, specifically the social distance and prior interaction, were the most important factors that influence pull request acceptance positively. Technical factors, specifically the number of comments in a pull request and the number of stars of a project, were the most important factors that influence the pull request acceptance negatively. These results were consistent with the findings of Tsay et al.'s work.

While the odds ratio of all the features varied slightly, we believe Tsay et al.'s results still stand. The change in some of the factors may be due to the selection of projects. Since we only included projects that have more than 250 closed or merged pull requests, it may have resulted in the selection of projects that have been active for a longer duration.

Given that Tsay et al. described how they obtained the GitHub repositories in great detail, it was relatively easy to fetch the exact same dataset, making the reproduction procedure seamless. In general, when working with a large dataset, we would encourage researchers to share their curated dataset publicly and outline how they obtained and filtered the dataset.

References

[1] J. Tsay, L. Dabbish, and J. Herbsleb, "Influence of social and technical factors for evaluating contribution in github," in *Proceedings of the 36th international conference on Software engineering.* ACM, 2014, pp. 356–366.

[2] N. Munaiah, S. Kroh, C. Cabrey, and M. Nagappan, "Curating github for engineered software projects," *Empirical Software Engineering,* vol. 22, no. 6, pp. 3219–3253, 2017.