Influence of social and technical factors for evaluating contribution in GitHub

ICSE 2014

Dyuti De
North Carolina State University
dde@ncsu.edu

ABSTRACT

Open source software is commonly portrayed as a meritocracy, where decisions are based solely on their technical merit. However, literature on open source suggests a complex social structure underlying the meritocracy. The authors present a study on open source software contribution in GitHub that focuses on the task of evaluating pull requests, which are one of the primary methods for contributing code in GitHub. Pull requests with many comments were much less likely to be accepted, moderated by the submitter's prior interaction in the project. Well-established projects were more conservative in accepting pull requests. It's found that project managers made use of information signaling both good technical contribution practices for a pull request and the strength of the social connection between the submitter and project manager when evaluating pull requests.

1 INTRODUCTION

The open source projects have complex socialization processes that need to be undertaken before accepting technical contributions. For evaluating the contributions many important factors such as technical norms, social connection, decision-making for highly discussed contribution, submitter's general community standing, submitter's status in project and project establishment has been explored. This helps understanding the underlying factors for acceptance of a pull request. [2]

2 METHOD

This paper tries to understand the factors affecting the acceptance of the contributions through the following research questions:

- **RQ1:** What is the effect of technical norms and discussions on pull requests.
- **RQ2:** How does the submitter's social connection, community and project standing help?
- **RQ3:** Does project's establishment has a role to play?

Pull requests have been used as a base unit to determine acceptance of contributions. It includes 659,501 pull requests across the 12,482 filtered projects and user information of 95,270 GitHub accounts. They have used the Github API to gather information on all issues and comments for each repository.

The associations have been measured with contribution in odds ratios. Using the pull request-level, submitter-level, and repository-level measures, they create a model that predicts the likelihood of pull request acceptance. They have fit a multi-level mixed effects logistic regression model to the data because the outcome variable (acceptance) is dichotomous and the dataset nested in multiple levels.

A logistic regression approach is chosen in order to better predict the dichotomous outcome variable. To ensure normality, each of the continuous variables in the model was log transformed and then centered such that the mean of each measure is 0 and standard deviation is 1.

Table 1: Descriptives of measures pre-transformation

Measure	mean	median	stdev	skew
Text Inclusion*	0.151	0.0	0.358	1.950
Commit Size	1456	25.00	27799	61.876
Files Changes	13.265	2.000	0	67.691
Social Distance*	0.096	0.000	0.295	2.740
Prior Interaction	200.583	22.000	8	8.184
Comments	2.664	1.000	6.656	19.198
Followers	35.972	7.000	2	22.965
Collaborator Status*	0.435	0.000	0.496	0.261
Repo Maturity**	2.104	1.956	1.188	0.568
Collaborators	20.203	8.000	42.808	6.063
Stars	1981	293	4095	2.977
Pull Req Acceptance*	0.723	1.00	0.447	-0.999

*Dichotomous variables **When data was collected, July 2017

3 ANALYSIS AND RESULTS

The research questions have been answered on the basis of model shown in Fig. 1. To account for the three-level nesting of the dataset from pull requests to users to repositories, they created a mixed model where the contribution measures are fixed effects and the unique user and repository intercepts are represented as random effects.

3.1 Pull Request level Measures

In terms of technical contribution norms, it is found that pull requests more consistent with community-wide pull request practices like inclusion of test cases and small commit sizes were more likely to be accepted. Code contributions that did not follow technical norms were less likely to be accepted, perhaps due to the higher assessment costs required by the project manager.

It is also found that social connections increase likelihood of contribution acceptance, even when controlling for compliance with technical contribution norms.

Foundations of Software Science, Fall 2018, Computer Science

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$9999.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

			Model I	Model II	Model III	Model IV
			Pull Request Level		Pull + Submitter Level	Pull + Submitter + Repo Level
	Factor	Variable	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio
Pull Request Level		(Intercept)	2.934 ***	2.898 ***	2.845 ***	3.925 ***
	Technical Contribution Norms (H1)	Test Inclusion	1.059 ***	1.023 *	1.114 ***	1.171 ***
		Commit Size	0.849 ***	0.834 ***	0.736 ***	0.738 ***
		Number of Files Changed	1.165 ***	1.152 ***	0.970 ***	0.927 ***
	Social Connection (H2)	Social Distance	1.345 ***	1.461 ***	3.636 ***	2.870 ***
		Prior Interaction	1.423 ***	1.362 ***	1.207 ***	1.356 ***
	Highly Discussed Contributions (H3)	Comments	0.481 ***	0.480 ***	0.414 ***	0.454 ***
	Decision-Making for Highly Discussed Contributions (H4)	*Test Inclusion x Comments*		1.057 ***	1.092 ***	1.106 ***
		Commit Size x Comments		1.101 ***	1.166 ***	1.169 ***
		Files Changed x Comments		1.017 ***	1.043 ***	1.035 ***
		Social Distance x Comments		0.806 ***	0.792 ***	0.796 ***
		Prior Interaction x Comments		1.106 ***	1.246 ***	1.142 ***
Submitter Level	Status in General Community (H5)	Followers			1.060 ***	1.181 ***
	Status in Project (H6)	Collaborator Status			3.904 ***	1.636 ***
Repo Level	Project Establishment (H7)	Repository Age				0.820 ***
		Collaborators				0.954 **
		Stars				0.648 ***
AIC:			633600	630879	506850	461077

Figure 1: Multi-level mixed effects logistic model for pull request acceptance

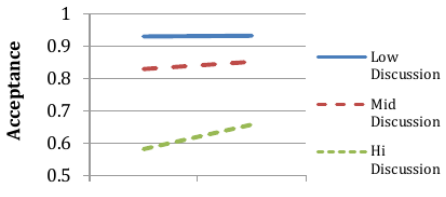

Figure 1. Interaction plot of test inclusion and contribution contention

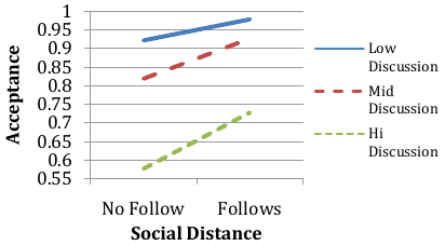

Figure 2. Interaction plot of social distance and contribution contention

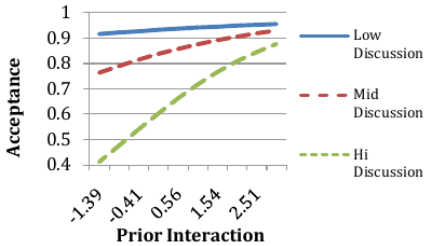

Figure 3. Interaction plot of prior interaction and contribution contention (prior interaction values are standardized)

3.2 Highly Discussed Contribution

In the model (Table 1) both technical contribution norms and social connection measures moderated the effect of discussion on contributions.

3.3 Submitter and Repository Level

Both the factors of 'followers' and 'collaborator status' is positively associated with the acceptance of pull-request. Well-established projects were more conservative when evaluating pull requests, perhaps due to audience pressures. In particular, the popularity of a repository has the strongest negative association out of the three.

4 LIMITATION

Some of the measures are not so robust to reverse-causality because of timing inherent in the pull request process. The social distance calculation assumed in the paper is not suffice. The OSS projects often maintain IRC or mailing lists over which contributors connect with the project managers [3], which remains dark to us. The future work should include the impact of social influence in the Github Repository. The research should also focus on getting a more vivid idea of domain knowledge of submitters. [1].

REFERENCES

[1] Kelly Blincoe, Jyoti Sheoran, Sean Goggins, Eva Petakovic, and Daniela Damian. 2016. Understanding the popular users: Following, affiliation influence and leadership on GitHub. *Information and Software Technology* 70 (2016), 30 – 39. https://doi.org/10.1016/j.infsof.2015.10.002

[2] James D. Herbsleb Jason Tsay, Laura A. Dabbish. 2014. Influence of social and technical factors for evaluating contribution in GitHub. In *ICSE'14, Hyderabad, India.* ACM.

[3] Georg von Krogh, Sebastian Spaeth, and Karim R Lakhani. 2003. Community, joining, and specialization in open source software innovation: a case study. *Research Policy* 32, 7 (2003), 1217 – 1241. https://doi.org/10.1016/S0048-7333(03)00050-7 Open Source Software Development.