SCIENCE CHINA
Information Sciences

• MOOP •

August 2024, Vol. 67, Iss. 8, 184101:1–184101:3
https://doi.org/10.1007/s11432-024-4090-0

Do community responses influence OSS contributor retention? A survival analysis

Zhe WANG[1], Zhixing LI[2*] & Yue YU[2]

[1]*School of Public Policy and Management, Tsinghua University, Beijing* 100871*, China;*
[2]*College of Computer, National University of Defense Technology, Changsha* 410073*, China*

Received 22 January 2024/Revised 18 March 2024/Accepted 7 June 2024/Published online 25 July 2024

Open source software (OSS) projects usually employ a community-based model, where geographically distributed developers from around the world collaborate online [1]. With the rise of social coding sites like GitHub, contributing to OSS projects has become easier and more accessible, motivating many developers to join the OSS community [2]. The continued submission of high-quality contributions from these community developers has led to numerous corrections, improvements, and innovations in many successful OSS projects [1, 2]. Therefore, attracting and retaining community contributors is crucial to the survival of OSS projects.

To assist OSS projects in retaining external community contributors, a large body of research has focused on understanding the factors that influence contributor retention in OSS projects [3–5]. However, the effects of community responses have not been well studied. OSS development is inherently collaborative and social, and prior studies have highlighted the importance of community responses in OSS collaboration. Despite this, there is a lack of studies quantitatively analyzing how factors related to community responses, such as whether a contributor received a reply or the sentiment of those replies, influence OSS contributor retention. Particularly, GitHub has introduced a social feature called reactions [6] (similar to emojis) that allows participants to easily express feelings. Investigating how these reactions are associated with contributor retention could offer valuable insights.

Moreover, individuals can participate in OSS projects in several different ways, such as submitting pull-requests (PRs), reporting issues, and posting discussion comments. In our preliminary examination, we examined and compared the survival rates of PR submitters (PRSs), issue reporters (IRs), and comment posters (CPs). Interestingly, we observed differences among the survival curves, and pairwise log-rank tests confirmed that these differences were statistically significant. It is important to explore to what extent these differences can be attributed to community responses.

To better understand OSS contributors' retention, we conducted a longitudinal survival analysis based on historical participation data collected from five OSS projects.

Methodology. Our analysis was conducted within the context of GitHub, one of the largest and most well-known OSS platforms. We focused on five GitHub projects: `Rails`, `Bootstrap`, `Ansible`, `Elasticsearch`, and `TensorFlow`. These projects were chosen because their popularity and activity levels provide representative and adequate data for our analysis. Moreover, these projects were written in five different programming languages and covered various application domains, adding diversity and representativeness to our data set.

For each project, we accessed the official GitHub API to gather data on contributors' activities from the creation date of that project until the examination date (October 19th, 2021). We excluded maintainers (i.e., core team members) and bots from our analysis to focus on peripheric participants in OSS communities. After removing bots' and maintainers' activities, we had a data set of 99015 contributors. We considered three types of participation activities in OSS projects: submitting PRs, reporting issues, and posting comments on others' PRs/issues. Our study concentrated on single-role contributors who performed a single type of activity and constitute the majority, including PRSs, IRs, and CPs. Multi-role contributors (e.g., contributors who not only reported issues but also submitted PRs), who accounted for only 20% of all contributors, were left for future work. To investigate how community responses affect OSS contributors' retention and how these effects vary among different types of contributors, we split the data into three subsamples based on contributor types and performed a survival analysis for each type.

In each survival analysis, we used the Cox proportional-hazards regression model [7]. This model is widely used in survival analysis to examine the effects of several observed variables on the hazard rate of a specified event (i.e., outcome variable). The outcome variable in our analysis is the occurrence of the leaving event: a contributor remains inactive in an OSS project for a long time. Following common practice [4], we defined 12 months of inactivity as the threshold for detecting a leaving event. Contributors were considered disengaged from a project in the current month if no activities were observed in the subsequent 12 time windows. Considering the right censorship, contributors whose last activity occurred less than 12 months prior to the end of our observation period were treated as still active. Additionally, 3.7% of contributors remained inactive for over 12 months

[*] Corresponding author (email: lizhixing15@nudt.edu.cn)

but later returned to make new contributions. Based on our definition of the leaving event, these contributors were considered exceptional cases and, therefore, excluded from our analysis.

The Cox regression model's input variables are the following. We aggregated the activities of each contributor by one-month time units and computed monthly values for all variables.

`replied`. This dummy variable indicates whether the contributor has received replies to her/his activities in the current month. The variable is set to true if the participant has received a reply from any of their historical activities.

`response_senti`. This variable represents the accumulated sentiment score computed over all responses received in the current month. Specifically, we used the state-of-the-art sentiment analysis tool Senti4SD, retrained from GitHub data, to analyze text sentiment [8]. For each reply text, Senti4SD categorizes sentiment polarity into negative, neutral, and positive, which we converted to numerical scores of −1, 0, and 1. If this variable exceeds 0, it indicates that the contributor has received more positive replies than negative ones in the current month.

`reaction`. This dummy variable shows whether the contributor received reaction responses in the current month. On GitHub, contributors can add eight different reactions to PRs, issues, and comments: `thumbsUp` (👍), `thumbsDown` (👎), `laugh` (😁), `confused` (😕), `heart` (❤️), `hooray` (🎉), `rocket` (🚀), and `eyes` (👀). Similar to the `replied` variable, this variable is set to true if the contributor has received a reaction from any one of their historical activities. We also included eight corresponding variables to represent the percentage of each type of reaction received in the current month.

Additionally, we included the following control factors:

`repo_age`. This variable indicates the time elapsed from when the project was hosted on GitHub to the current month (measured in months).

`n_stars`. This variable denotes the number of stars that the project has received until the current month.

`n_forks`. This variable represents the number of forked repositories of the project until the current month.

`n_commits`. This variable indicates the number of commits made to the project's source code repository until the current month.

`n_tasks`. This variable measures the number of new issues and PRs submitted to the project during the current month.

`n_contrs`. This variable indicates the number of active contributors to the project during the current month.

`act_counts`. This variable measures the number of activities that the contributor has performed during the current month.

`act_desc`. Generally, contributors provide a textual description when reporting an issue or submitting a PR to OSS projects. This variable indicates the average length of issue/PR description or comments created by the contributor in the current month.

When building the models, we considered multicollinearity between the variables and removed variables, as needed, using the variance inflation factor (VIF) indicator, compared to the maximum of 5, which was commonly adopted in previous studies. In our case, the variables `n_forks` and `n_commits` were removed owing to multicollinearity. Among the remaining variables, the highest VIF value was 4.13, indicating that multi-collinearity was not a concern.

Results. Table 1 presents summaries of our three Cox regression models, which reveal the influence of various variables on OSS contributors' sustained participation. Model-1 (N = 11086 contributors), Model-2 (N = 38792 contributors), and Model-3 (N = 45470 contributors) were built based on the participation activity data of PRSs, IRs, and CPs, respectively. The results are reported in the form of hazard ratios (HRs). An HR above 1 indicates a positive association with the event probability (i.e., leaving the project in our context), while an HR below 1 indicates a negative association.

From the table, we observe that all control variables have a significant effect, with four of them presenting consistent effects among all three models. For instance, the variables `repo_age` and `n_stars` have HR values below 1, indicating that as the project becomes more mature and popular, contributors, regardless of their role, are more likely to remain involved. By contrast, the variables `act_counts` and `act_descs` have HR values above 1, indicating that contributors have a higher probability of leaving the project when they are engaged in heavier and more complex tasks. Interestingly, the variables `n_tasks` and `n_contors` present significant but inconsistent effects across the three models. For the variable `n_tasks`, we observe that when there are more newly submitted tasks in a project, both PRSs (HR = 0.739 in Model-1) and IRs (HR = 0.894 in Model-2) are less likely to leave the project. On the contrary, CPs have a higher chance of leaving, as indicated by an HR value greater than 1 (i.e., 1.494) in Model-3. As for the variable `n_contors`, we find that when there are more active contributors in the project, both PRSs (HR = 1.505 in Model-1) and IRs (HR = 1.183 in Model-2) have a higher probability of leaving the project. However, CPs are more likely to stay in the project (HR = 0.774 in Model-3).

For variables related to community responses, the variable `replied` presents significant but different effects across the three models. Specifically, the HR values in Model-1 (HR = 1.156), Model-2 (HR = 1.206), and Model-3 (HR = 0.921) indicate that CPs are more likely to retain when receiving responses, while PRSs and IRs have a higher chance of leaving the project upon receiving responses. Interestingly, the HR values of the variable `replied X response_senti` in Model-1 (HR = 0.939) and Model-2 (HR = 0.938) show that a higher percentage of positive feedback in the received responses would decrease the possibility of PRSs and IRs leaving.

The significant effects and HR values of the variable `reaction` in Model-1 (HR = 0.719) and Model-2 (HR = 0.853) show that responding with reactions helps to retain RPSs and IRs. However, CPs are not significantly affected. As for the specific type of reactions, only the reactions 😁 and 🎉 did not present any significant effect in all three models. The variable `reaction X thumbsUp` presents a significant, negative effect in Model-2 (HR = 1.090), indicating that IR retention would be negatively affected by the reaction 👍. Although this influence is slight, it is the exact opposite of what we would have expected, i.e., the reaction 👍 expresses a positive opinion, and contributors receiving this reaction would be more willing to stay in the project. As expected, the variable `reaction X thumbsDown` has a significant, negative effect in Model-1 (HR = 1.516) and Model-2 (HR = 1.237). This means PRSs and IRs have a higher chance of leaving the project when they receive the reaction 👎, which represents responders' negative opinions. How-

Table 1 Results of the three Cox proportional-hazards regression models[a)]

	Variables	Model-1 (PRS)		Model-2 (IR)		Model-3 (CP)	
		HR	SE	HR	SE	HR	SE
Controls	repo_age	0.848***	0.013	0.848***	0.007	0.906***	0.006
	n_stars	0.868***	0.012	0.826***	0.007	0.809***	0.007
	n_tasks	0.739***	0.041	0.894***	0.023	1.494***	0.013
	n_contors	1.505***	0.043	1.183***	0.024	0.774***	0.016
	act_counts	1.272***	0.027	1.520***	0.021	1.143***	0.012
	act_descs	1.033***	0.005	1.147***	0.004	1.151***	0.004
Community responses	replied	1.156**	0.026	1.206**	0.017	0.921**	0.015
	replied X response_senti	0.939***	0.024	0.938***	0.013	0.899***	0.015
	reaction	0.719***	0.700	0.853***	0.185	0.650	0.111
	reaction X thumbsUp (👍)	1.020	0.063	1.090**	0.034	0.963	0.022
	reaction X thumbsDown (👎)	1.516***	0.118	1.237***	0.073	1.030	0.042
	reaction X eyes (👀)	0.681	0.919	0.910	0.124	0.803*	0.087
	reaction X confused (😕)	1.944***	0.353	1.084	0.102	0.974	0.067
	reaction X laugh (😁)	0.867	0.248	1.155	0.103	1.037	0.062
	reaction X hooray (🎉)	0.898	0.112	0.898	0.095	0.952	0.044
	reaction X rocket (🚀)	0.515*	0.266	0.653*	0.217	0.817	0.114
	reaction X heart (❤️)	0.820*	0.089	1.053	0.072	0.912*	0.044
	R^2	0.07		0.12		0.17	

a) *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.
Abbreviations: HR – hazard ratio; SE – standard error; PRS – PR submitters; IR – issue reporters; CP – comment posters.

ever, such influence was not observed for CPs. The variable **reaction X eyes** only presents a significant effect in Model-3, with an HR value of (HR $= 0.803$), indicating that CPs are more likely to stay in the project when receiving the reaction 👀. Surprisingly, the HR value (HR $= 1.944$) of the variable **reaction X confused** in Model-1 indicates that the probability of leaving the project for PRSs would increase by 94.4% ($(1.944 - 1) \times 100\%$), if they received the reaction 😕. In both Model-1 and Model-2, the variable **reaction X rocket** presents a significant, positive effect, meaning that PRSs and IRs are more likely to stay in the project when they receive the reaction 🚀. As for the variable **reaction X heart**, we observe a significant, positive effect in both Model-1 and Model-3, indicating that PRSs and CPs would have a lower chance of leaving when they received the reaction ❤️.

Conclusion and implication. Our survival analysis provides evidence that positive responses from the community help retain contributors in OSS projects, and PRS are more sensitive about reactions than IRs and CPs. Particularly, reactions 👎 and 😕 would significantly increase PRS' possibility of leaving the project, while reactions 🚀 and ❤️ positively impact their willingness to stay.

Given the importance of retaining OSS contributors, we suggest the following: (1) OSS community members should be cautious when making comments or reactions with negative implications to avoid causing disengagement; (2) OSS platforms can design tools to automatically detect discussion threads with an excessive number of negative responses and call for necessary intervention from project maintainers.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2021ZD0112904) and Key Laboratory Fund Project (Grant No. 2022-KJWPDL-04).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Raymond E. The cathedral and the bazaar. Know Techn Pol, 1999, 12: 23–49

2 Wang H M, Yu Y, Wang T, et al. Crowd intelligence paradigm: a new paradigm shift in software development (in Chinese). Sci Sin Inform, 2023, 53: 1490–1502

3 Zhou M, Mockus A. What make long term contributors: willingness and opportunity in OSS community. In: Proceedings of the 34th International Conference on Software Engineering (ICSE), 2012. 518–528

4 Qiu H S, Nolte A, Brown A, et al. Going farther together: the impact of social capital on sustained participation in open source. In: Proceedings of the 41st International Conference on Software Engineering (ICSE), 2019. 688–699

5 Lin B, Robles G, Serebrenik A. Developer turnover in global, industrial open source projects: insights from applying survival analysis. In: Proceedings of the 12th International Conference on Global Software Engineering (ICGSE), 2017. 66–75

6 Borges H, Brito R, Valente M T. Beyond textual issues: understanding the usage and impact of github reactions. In: Proceedings of the XXXIII Brazilian Symposium on Software Engineering, 2019. 397–406

7 Kleinbaum D G, Klein M. Survival Analysis. Berlin: Springer, 1996

8 Calefato F, Lanubile F, Maiorano F, et al. Sentiment polarity detection for software development. In: Proceedings of the 40th International Conference on Software Engineering (ICSE), 2018