Towards Unraveling Developers Communities in Stack Overflow and Reddit

Keywords: User behavior, Q&A social platforms, Developer communities, Hypergraphs

Extended Abstract

Introduction. Today, coding skills are among the most required competencies worldwide, often also for non-computer scientists. Because of this trend, community contribution-based, question-and-answer (Q&A) platforms became prominent for finding the proper solution to programming issues [2]. Stack Overflow[1] (SO) has been the most popular platform for technical-related questions for years. Still, recently, some programming-related subreddits of Reddit[2] have become a reference place for technical discussions. Currently, few works compare developer communities across similar structured social Q&A sites, either focusing on learning practice or specific topics. In this work, we investigate the developers' behavior and community formation from a broader perspective by performing a longitudinal study on users' posting activity and by originally modeling their high-order interaction patterns via hypergraphs [1] (which allowed us to link users that have interacted with the same question, even though they did not directly answer each other - see Figure 1).

Preliminary results. In this study, we focused on the top 20 most used programming languages (cf. SO survey[3]) and collected SO data using Kaggle, while Reddit data via the Pushshift API. The final data set covers a two-year period from 2020/09/24 to 2022/09/24. To assess the evolution of the posting patterns and interactions of SO and Reddit users over the two years, we split our dataset into eight snapshots, each 3-months long. For each period, we evaluated both quantitative and structural information.

· *SO and Reddit users profoundly differ in their activity patterns.* These differences are clearly evident from the data we collected in terms of *(i)* the number of users posting and/or answering questions (see Figure 2), and *(ii)* the length (number of comments) and time-span of each conversation (see Figure 3). The first interesting outcome is the steeply decreasing trend in the number of users and questions/comments on SO, potentially due to the increasing knowledgeability of the developers and a large database of existing questions. In contrast, the number of Reddit users (i.e., Reddittors) tends to increase gently over time, possibly indicating the presence of a persistent developer community on the platform. Regarding conversation-specific characteristics, SO conversations are characterized by few comments, with most questions receiving only one answer. Longer conversations extend up to 20 comments on average. On the contrary, Reddit discussions appear fairly longer compared to SO, with half of the conversations made up of $4/5$ comments. Generally, SO questions tend to be answered quickly, potentially due to the platform's gamified mechanism. However, some discussions can last over a year. On Reddit, half of the conversations last up to 15-17 hours, and another portion extends over more than one day. The longest Reddit discussions span no more than three months, and the reason for this is unclear and may require further research.

· *SO and Reddit users engage in discussions about different programming languages on the two platforms.* Figure 4 depicts this information, considering generic-purpose and domain-specific

[1] https://stackoverflow.com/
[2] https://www.reddit.com/
[3] https://survey.stackoverflow.co/2022/

languages independently. We can note that Python-related questions dominate on SO, followed by C#-related queries, while on Reddit, Python, Rust, and Clojure-related questions are equally popular. Generally, no other technology sensibly prevails over the others, even though they follow different trends on both platforms. We can observe a similar situation when looking at the proportion of queries about domain-specific languages. Overall these results suggest that SO is a referring platform for asking questions related to commonly used programming languages, such as Python and Javascript. On the contrary, Reddit appears to be preferred to discuss recently rising programming languages, like Rust. Once again, this trend may be due to the inherent nature of the Reddit website, where users can either discuss or ask for help on a technical topic.

· *The Reddit discussion model seems to represent a fertile field for the growth of more supportive and long-lasting communities than SO.* The label propagation community detection algorithm ran on the user-question interaction hypergraphs highlighted a considerable number of communities discovered in both Q&A platforms (see Table 1), with most of them composed of very tiny groups of users. In our analysis, we focused our attention on the op 5% largest communities on both platforms, characterizing them according to the most common question tag. As expected, the biggest communities are represented by trending languages like Python and Javascript for SO, and Python, Rust, and Go for Reddit. However, some languages like Clojure and C# have low user representation in some communities, which may suggest a lack of interaction for a specific topic or a few committed users who heavily discuss a given language. The study of the communities' evolution (see Figure 5) suggests that most SO users are interested in receiving feedback for solving their problems rather than contributing and building a community around specific topics. However, this outcome may change if users are observed over more extended periods. In contrast, on Reddit, a more significant portion of users consistently contribute to the platform, reflecting a greater sense of bonding among community members in addition to task-specific discussions.

Conclusion and Future Work. This preliminary study highlighted significant differences in how SO and Reddit are utilized in terms of conversation length, duration, trending technologies, as well as community formation, and evolution. These differences arise from the platforms' design choices and social mechanics, ultimately influencing their popularity and ability to sustain an active community of contributors. Future work will encompass two main directions. On one side, we will examine a more extensive user base to grasp user behavior changes and community evolution across different time periods. On the other side, we will conduct a more detailed characterization of each community structurally and semantically (based on the content of the text to identify discussion subtopics). We further plan to analyze discussions related to AI-generated code via GitHub Copilot[4] and ChatGPT [5] to unravel how these trending tools are replacing standard Q&A websites in daily use.

References

[1] F. Battiston et al. "Networks beyond pairwise interactions: Structure and dynamics". In: *Physics Reports* 874 (2020), pp. 1–92.

[2] G. Ndukwe et al. "Perceptions on the Utility of Community Question and Answer Websites Like Stack Overflow to Software Developers". In: *IEEE TSE* (2022), pp. 1–13.

[4] https://github.com/features/copilot

[5] https://chat.openai.com/chat

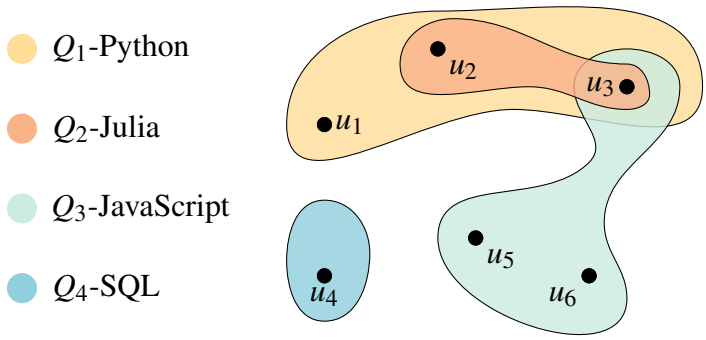


Figure 1: An example of user-question interaction hypergraph, where vertices represent SO or Reddit users and hyperedges their indirect interactions when they answer the same question or submission. Each hyperedge is labeled with the topic of the question. For instance, the users u_1, u_2, and u_3 reply to the same question Q_1 tagged as Python-related but do not necessarily comment on each other's posts.

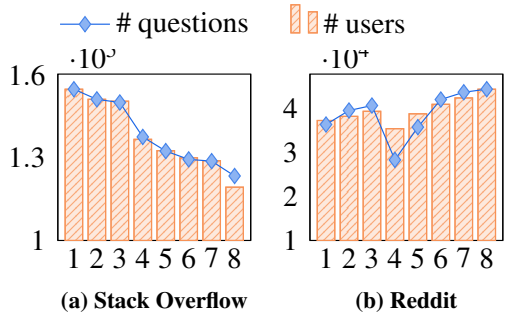
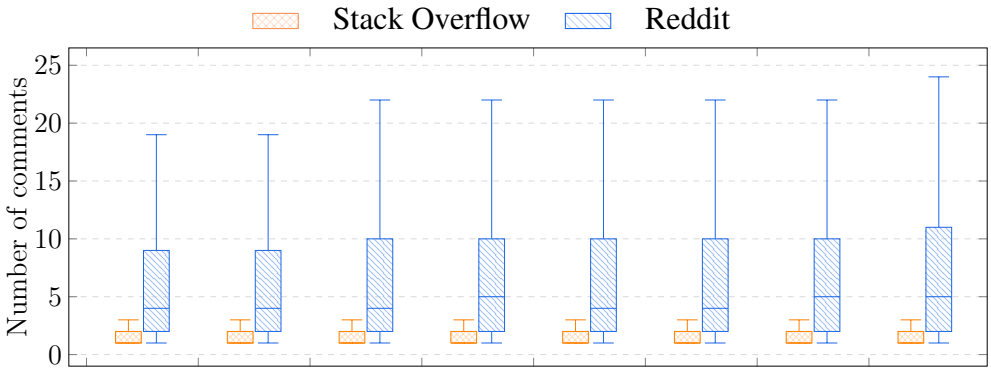
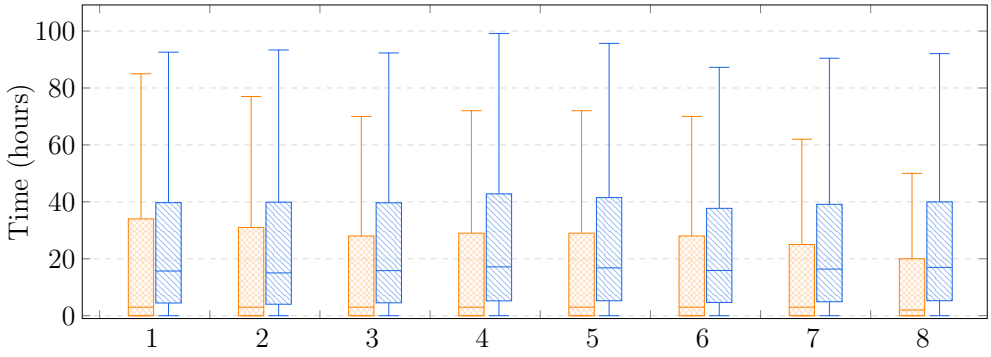


Figure 2: Number of users that have asked or answered at least one question plotted against the overall number of questions per trimester.



(a) Distribution of the length of conversations.



(b) Time span of conversations.

Figure 3: Comparison of the number of answers and durability of conversations in Stack Overflow and Reddit per trimester.

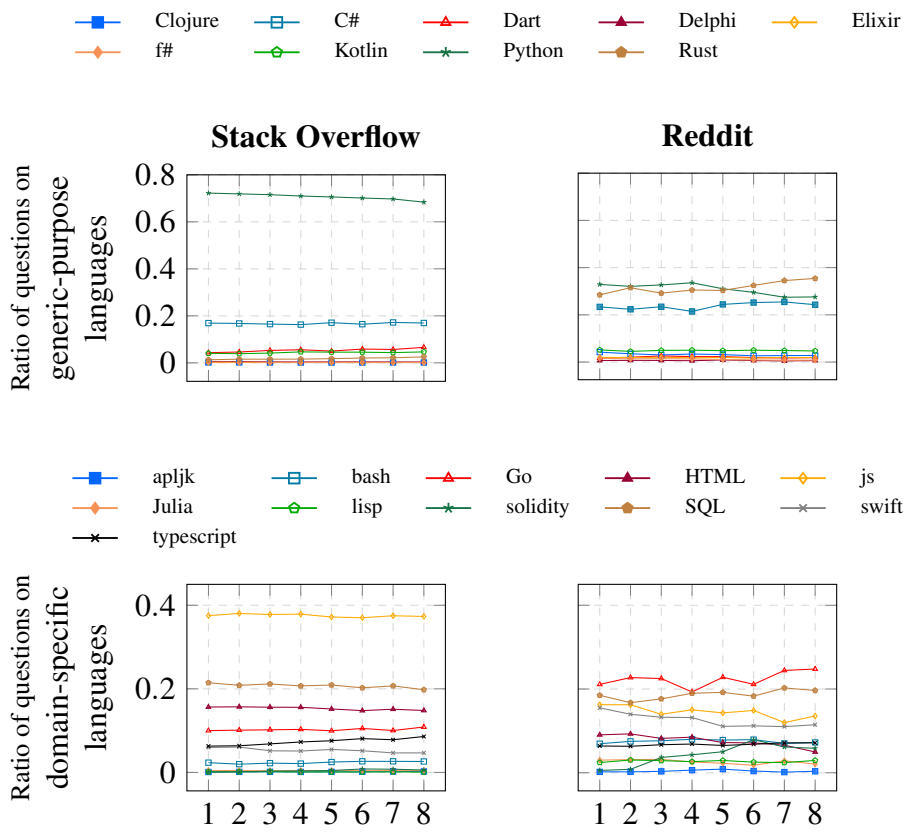
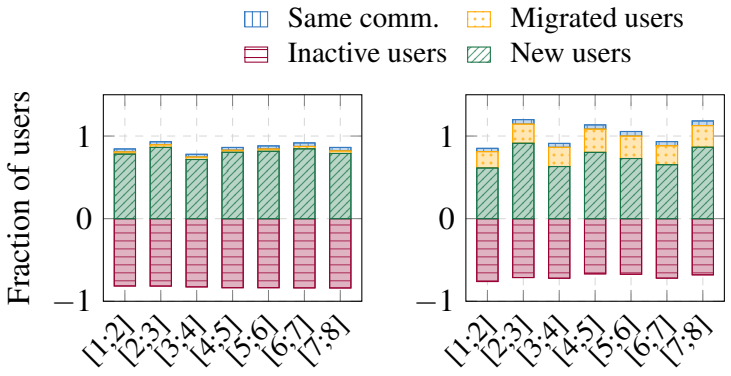


Figure 4: Proportion of the number of each language-related questions in Stack Overflow and Reddit per trimester.

Trimester		1	2	3	4	5	6	7	8
C#	SO	2,96	3,32	2,95	2,89	3,52	2,66	4,24	3,48
	Reddit	14,52	15,77	11,33	12,50	14,12	14,34	15,55	13,62
Clojure	SO	0,00	0,02	0,04	0,00	0,06	0,00	0,00	0,04
	Reddit	1,10	0,79	0,63	1,02	0,33	0,68	0,61	0,65
Go	SO	0,31	0,36	0,20	0,44	0,43	0,34	0,55	0,27
	Reddit	7,74	7,50	6,88	9,08	9,67	8,33	12,30	7,02
Javascript	SO	25,29	26,72	24,02	27,64	28,84	27,81	24,22	25,29
	Reddit	7,68	9,42	9,95	7,84	7,93	8,47	7,14	6,26
Python	SO	61,13	58,36	63,51	58,76	56,51	57,24	58,78	59,22
	Reddit	33,71	36,21	28,37	32,34	32,84	25,70	31,45	30,22
Rust	SO	0,20	0,30	0,28	0,23	0,20	0,39	0,29	0,40
	Reddit	23,94	18,02	28,71	26,64	26,85	28,73	20,44	30,55
SQL	SO	4,01	3,88	3,86	4,01	4,93	4,56	5,86	5,51
	Reddit	1,66	2,29	2,16	2,14	1,65	4,12	4,61	3,04

Table 1: Percentage of users per programming language within the top 5% biggest communities.



(a) Stack Overflow. (b) Reddit.

Figure 5: Community evolution evaluated considering the user behavior in terms of whether they stay active in the same community (i.e., continue to ask or answer questions related to their current community), migrate to another community (i.e., begin to consistently ask/answer questions in another community), go inactive (i.e., stop asking/answering questions anywhere on the platform of interest even though they can still lurk on the platform, or (re-)join the community (i.e., newly registered users or users that become active again after having been inactive the previous trimester(s)). The figure shows the percentage of these four classes of users evaluated pairwise over consecutive trimesters. The values shown are averaged over all communities.