The Age of Snippet Programming: Toward Understanding Developer Communities in Stack Overflow and Reddit

The Age of Snippet Programming: Toward Understanding Developer Communities in Stack Overflow and Reddit

Alessia Antelmi[*]
aantelmi@unisa.it
Università degli Studi di Salerno

Gennaro Cordasco
gennaro.cordasco@unicampania.it
Università della Campania “Luigi Vanvitelli”

Daniele De Vinco[*]
ddevinco@unisa.it
Università degli Studi di Salerno

Carmine Spagnuolo
cspagnuolo@unisa.it
Università degli Studi di Salerno

ABSTRACT

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 all programming issues. Stack Overflow has been the most popular platform for technical-related questions for years. Still, recently, some programming-related subreddits of Reddit have become a standing stone for questions and discussions. This work investigates the developers' behavior and community formation around the twenty most popular programming languages. We examined two consecutive years of programming-related questions from Stack Overflow and Reddit, performing a longitudinal study on users' posting activity and their high-order interaction patterns abstracted via hypergraphs. Our analysis highlighted crucial differences in how these Q&A platforms are utilized by their users. In line with previous literature, it emphasized the constant decline of Stack Overflow in favor of more community-friendly platforms, such as Reddit, which has been growing rapidly lately.

CCS CONCEPTS

• **Human-centered computing → Social media**; • **Mathematics of computing → Hypergraphs**.

KEYWORDS

User behavior, Q&A social platforms, Developer communities, Stack Overflow, Reddit, Hypergraphs, Network Analysis

ACM Reference Format:
Alessia Antelmi, Gennaro Cordasco, Daniele De Vinco, and Carmine Spagnuolo. 2023. The Age of Snippet Programming: Toward Understanding Developer Communities in Stack Overflow and Reddit. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3543873.3587673

[*] Corresponding authors.

WWW '23 Companion, April 30-May 4, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587673

1 INTRODUCTION

Programming enthusiasts love spending their time coding solutions to complex challenges or printing a trivial "Hello World" in the newest, most popular programming languages. In most cases, they rely on Questions-and-Answers (Q&A) websites to acquire knowledge, solve problems, seek snippets of code for reuse, improve their own code, and discuss technical concepts [23]. The social mechanics of these platforms shape the formation of topic-specific communities and play a critical role in how information is shared within them. Still, these platforms' similar structures and functionalities can present a challenge when looking for the proper place to seek help. In this paper, we are interested in unraveling the developers' behavior and community formation around the most used programming languages in Q&A platforms. Specifically, our first objective was to examine whether and, if so, how users' activity patterns differ from one website to another in terms of the length and duration of conversations, as well as the platform-specific trending topics. Our second intent was to investigate the community structure of different platforms, characterize their trending technologies, and evaluate their persistence or evanescence over time.

This study represents a preliminary study toward answering these questions by gathering data about the twenty most used programming languages in two popular Q&A platforms, i.e., Stack Overflow and Reddit. The main contributions of our work can be summarized as follows:

- Stack Overflow and Reddit profoundly differ in how their users communicate in terms of the length and duration of conversations. These patterns naturally reflect the platforms' different intents in sharing knowledge.
- The most discussed languages in Stack Overflow do not correspond to those in Reddit. As in the previous case, this outcome may be due to the inherent nature of both Q&A platforms, as users prefer Reddit over Stack Overflow to discuss newer and trending languages.
- Both platforms exhibit the typical inverse pyramid of contribution [26], where most users contribute poorly to creating new content while only a few produce most of it.
- Overall, the analysis of the evolution of Stack Overflow communities suggests the falling out of the Q&A website in favor of other platforms. At the same time, it highlights Reddit's rise in the programming domain.

Our work can be framed under the general research direction of understanding the dynamics of participation in Q&A platforms to encourage valuable user engagement and improve the worth of crowd-sourced knowledge. Nowadays, these objectives are becoming even more critical since we are entering the era of AI-generated code thanks to tools like GitHub Copilot [13] or ChatGPT [27]. Despite their help, developers must still have some coding skills to fully exploit their potential; in this context, supporting communities will continue playing a crucial role as learning environments [17].

2 BACKGROUND

This section introduces the reader to the Q&A online platforms analyzed in this work and the main notions of hypergraphs.

Reddit and Stack Overflow. Q&A online social platforms lay their foundations in the wisdom of crowds [38]: everyone can ask and answer questions, thus, contributing to increasing community knowledge. Currently, there is a myriad of Q&A websites, which share some common features, such as the use of reputation/points awarded for providing good questions and answers. In this work, we consider two of the most popular social Q&A platforms where users can ask programming language-related questions: Stack Overflow [28] and Reddit [31].

Stack Overflow is one of the most popular Q&A platforms specifically designed for developers. Users can submit a programming-related question, attach tags to it, and receive answers from software developer communities. According to its guidelines, a core characteristic of Stack Overflow is that users should avoid opinion-based questions that could generate discussions rather than answers. Reddit is one of the most popular social network websites focused on news aggregation and discussion, mainly in the form of question and answer. The key element of this platform is its organization in subreddits, which represent online communities centered around a specific topic, like technology, politics, or sport. Although quite different in their intent, Stack Overflow and Reddit share some similarities: both *(i)* are community-driven, and *(ii)* live on the content generated by their users, *(iii)* use a scoring system to assess user reputation, and *(iv)* allow to filter their content by categories easily.

Hypergraphs. A hypergraph is a generalization of a graph where a (hyper)edge allows the connection of an arbitrary number of nodes [10]. Such structures are the natural representation of a broad range of systems where group-structured relationships exist among their interacting parts. For instance, hypergraphs can easily abstract social systems where individuals interact in groups of arbitrary size, as in the case of a co-authorship collaboration network, where a hyperedge may represent an article and link together all authors (nodes) having collaborated on it. Similar situations, characterized by high-order interactions, also exist in biology, ecology, and neuroscience [7]. The powerful expressiveness of hypergraphs has a few drawbacks: dealing with the complexity of such data structures and the need for appropriate tools and algorithms for their study. For this reason, hypergraphs have been little used in literature in favor of their graph counterpart. Recently, the trend of using hypergraphs to graph representations is drifting, thanks to an increasing number of systematic studies demonstrating how the transformation of a hypergraph to a classical graph either leads to an inevitable loss of information or creates a large number of extra nodes/edges that increases space and time requirements in downstream graph analytic tasks [25, 41].

3 RELATED WORK

Q&A websites have become a valuable source of shared knowledge related to a wide variety of fields [32]. In this section, we specifically review literature focused on examining developer communities on two Q&A platforms, Stack Overflow and Reddit.

Stack Overflow has been extensively studied in the literature, including tasks such as predicting question tags, detecting duplicate questions, examining the usage of code snippets, and analyzing user behavior and community evolution [1, 22]. In particular, under this last perspective, some relevant works focus on characterizing the social evolution of user communities over a significant temporal period [9, 23] and detecting patterns of interest change [11]. Specifically, Blanco et al. [9] explore the social evolution experienced by the Java developer community over 10 years, assessing the evolution of topics, user reputation, and cross-references of internal contents and external sources to improve the user experience by, for instance, reducing the number of questions with no answers. Moutidis et al. [23] examine the user evolution from a larger angle by considering all questions, answers, votes, and tags from Stack Overflow between 2008 and 2020 to evaluate trends in domain-related technologies and user persistence within the associated communities. Similarly, Fu et al. [11] tackle the problem of detecting patterns of interest change, but from a microscopic perspective, by quantifying how individuals shift their interest focus over time based on question tags.

Reddit has become a golden pot for researchers thanks to the openness, richness, and quality of its data [21]. In the context of developer communities, this platform has been analyzed to gather insights about how users contribute to global knowledge and learn from others in relation to computer science-specific issues. For instance, Waller et al. [39] explore users' activity diversity, characterizing whether they are specialists in some topics or contribute to broader knowledge. Lu et al. [18] focus on communications of game developers about their game development experiences, processes, and practice. Both Aniche et al. [2] and Zang et al. [42] compare Reddit with HackerNews [14] via developer surveys. While the former focuses on understanding what motivates developers from both communities to contribute and what kind of content is shared, the latter targets identifying barriers to the adoption of Rust. More recently, Liu et al. [17] explored how Reddit has been utilized by online learners to learn programming-related topics.

To our knowledge, few works currently compare developer communities across similar structured social Q&A sites. An example is the work from Sengupta [33], which explores differences in the discourse patterns in StackOverflow and Reddit (r/Askprogramming), specifically focusing on the learning practices these collectives support and scaffold. Wu et al. [40] examine similarities and discrepancies of the Reddit, Stack Overflow, and Stack Exchange cybersecurity communities with the purpose of better understanding how and why users choose one platform over the others when their needs are identical. Another example is the work of Luo et al. [20], where the authors analyze practitioners' opinions about low-code development on Stack Overflow and Reddit.

Our study differs from current literature along two axes, i.e., the targeted user base and the use of hypergraphs. Our aim is to analyze the Stack Overflow and Reddit communities' evolution around programming languages from a quantitative and structural perspective. In particular, we exploit hypergraphs to mine the high-order relationships between users and questions.

4 METHODOLOGY

This section concisely describes the methodology followed, starting from the data collection and preparation process to the analysis of users’ activity and interaction patterns. All code used to analyze the data and the results obtained are available on GitHub [3] while the datasets are available on Zenodo [4].

Data collection. In this study, we restrict our analysis to programming language-related questions. Specifically, we picked the top 20 most used programming languages in the Stack Overflow survey [36] to define the data filtering criteria. We queried Stack Overflow data from Kaggle [16] while we used the Pushshift API [30] to retrieve Reddit data. For both platforms, we collected: *(i)* the identifier of the comment, *(ii)* the identifier of the Stack Overflow question/Reddit submission, *(iii)* the id of the comment/question author, (iv) the date of creation, *(v)* the date of last interaction with the question, and *(vi)* the topic of the question/comment (i.e., Stack Overflow tag or subreddit name). After the data collection process, we anonymized all ids. The resulting data set spans over two years, starting from 2020/09/24 to 2022/09/24.

Data Analysis To assess the evolution of the posting patterns and interactions of Stack Overflow and Reddit users over the two years, we split our data set into eight snapshots, each 3-months long. For each period, we evaluated both quantitative and structural information.

Mining users’ activity patterns. As a first insight into each platform’s usage, we evaluated: *(i)* the number of unique users, *(ii)* the number of questions/comments, *(iii)* the distribution of the length and *(iv)* time span of conversations. We then focused on examining the popularity of the 20 languages in terms of the number of questions/submissions related to the specific topic.

Mining users’ interaction patterns. A set of users commenting and discussing the same question can be seen as part of the same many-to-many (or high-order) relation. Such type of relationship does not imply a direct pairwise interaction among each pair of users, but it only indicates they have commented on the same topic. In this work, we leverage hypergraphs to model the underlying question-answers interaction network, which allows us to highlight the existing relationships between users and questions rather than between users. The rationale behind this network creation approach is to link users that have interacted with the same question, even though they did not directly answer each other. In this way, we are able to cluster together groups of users with the same interests even if they have never interacted directly.

To address the above task, we built one hypergraph per trimester, whose vertices represent Stack Overflow and Reddit users, and hyperedges their indirect interactions when they answer the same question or submission. After labeling each hyperedge with the topic of the question, we ran the Label Propagation community detection algorithm [5] on each hypergraph snapshot and characterized each community based on the most prominent tag within it. For the bigger communities, we further investigated their persistence in terms of the users that constitute them over time.

5 RESULTS AND DISCUSSION

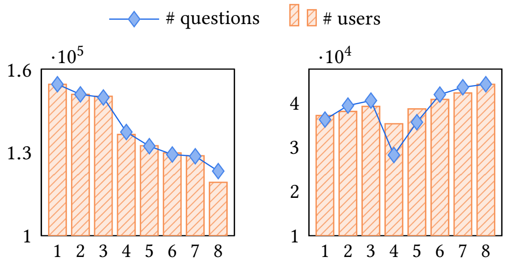
This section presents the results of our analyses, first focusing on a quantitative characterization of the platforms’ usage and language trends, then discussing user communities and their evolution.

5.1 Mining users’ activity patterns

It may be argued that among the first indicators of a platform’s liveliness, we can name the number of users actively engaging with others as well as the number of their activities on the platform itself. In the realm of Q&A websites, looking at how users connect with each other may guide toward a deeper comprehension of how users perceive and actually use the platform.

5.1.1 Usage of Stack Overflow and Reddit over time. Although Stack Overflow and Reddit represent two examples of Q&A online platforms, they profoundly differ in the way they are supposed to be used and in the intended user base (see Section 2). These differences are clearly evident from the data we collected in terms of *(i)* the number of users posting and/or answering questions about the selected programming languages, and *(ii)* the length (number of comments) and time-span of each conversation (i.e., *questions* for Stack Overflow, *submissions* for Reddit).

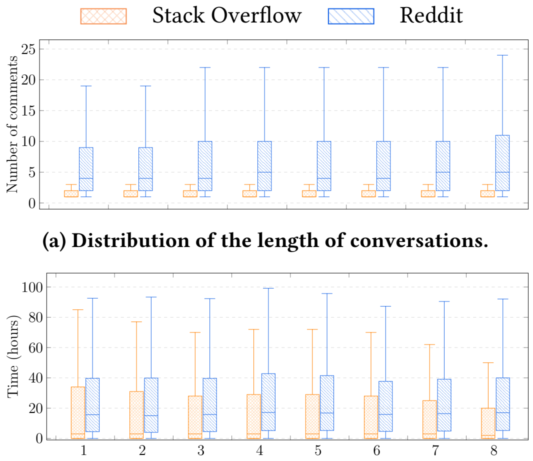
Figure 1 shows, for each trimester, the number of users that have asked or answered at least one question plotted against the overall number of questions. The first interesting outcome is the steeply decreasing trend in the number of users and questions/comments on Stack Overflow (see Figure 2a). This phenomenon may be justified by the increasing knowledgeability of the developers, which could be more skilled in tuning existing answers to their problems [9], as well as by the ever-increasing database of questions which makes it easier for someone to find another user who had already experienced the same issue. Another reason for this negative trend could also be sought in Stack Overflow’s increasing churn rate (i.e., users who leave the platform after an initial use [24]), possibly due to the constant evolution of the proportions of content quality and user types (in relation to the posting and answering behavior) over time [34]. Generally, the number of questions is comparable to the number of users, hence, suggesting a ratio of 1:1 between users and questions. In contrast, the Reddit data set exhibits a rather different perspective (see Figure 2b). In this case, the number of Reddit users (i.e., Redditors) tends to increase gently over the two years of



(a) Stack Overflow

(b) Reddit

Figure 1: Number of users and questions per trimester.



(a) Distribution of the length of conversations.

(b) Life span of conversations.

Figure 2: Comparison of the number of answers and durability of conversations in Stack Overflow and Reddit.

observation, possibly indicating the presence of a persistent developer community on the platform. Reddit's characteristic of being a domain-agnostic discussion website explains why the number of Redditors is one order of magnitude less than the number of Stack Overflow users in our data set. In the case of Reddit, we are only considering a very small subset of its user base (52 million daily active users on average [37]); in the case of Stack Overflow, we are naturally including a bigger slice of the overall population considering that the platform specifically addresses developers' questions. As before, the number of questions is proportional to the number of Redditors, even though questions slightly outnumber users. An exception is made for the fourth and fifth trimesters, where the situation is reversed.

Figure 2 reveals two very different patterns of interaction across the two platforms regarding the length (see Figure 2a) and the time span (see Figure 2b) of conversations. Stack Overflow 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 Stack Overflow, with half of the conversations made up of 4/5 comments. Reddit's underlying nature as a discussion forum is reflected in the long tail of the distribution, where 10% of the submissions receive 20/25 replies on average, with some outliers reaching a mean of 600 comments. Regarding the lifespan of conversations, it is not surprising that Stack Overflow questions tend to be answered in less than 2/3 hours. A possible motivation for this behavior may be sought in the platform's gamified mechanism, which stimulates volunteers' contributions by maximizing the chances of winning gamification rewards for fast answers [19]. Still, many Stack Overflow conversations span over one day or more, with some discussions still active after more than one year. In the case of Reddit, half of the conversations tend to last up to 15-17 hours, while another consistent portion extends over more than one day. However, conversely to the Q&A computer science platform, the longest Reddit discussions span no more than three months. The reason for this result may be the focus of further research.

5.1.2 *Technology trends.*

Another interesting perspective under which Stack Overflow and Reddit profoundly differ is the popularity (in terms of the number of questions) of each programming language. Figure 3 depicts this information, considering general-purpose and domain-specific languages independently to make it easier to read the plots. The y-axis reports the ratio of questions about a given language within the category of interest to enable comparison between the two platforms. The first noticeable outcome from this analysis regards the prevalence of Python-related questions on Stack Overflow (~80%). Together with C#-related queries (~20%), these constitute almost the totality of all raised issues about general-purpose languages. In the case of Reddit, Python, Rust, and Clojure-related questions share the stage with around 30/40% of questions each. 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. JavaScript-related questions dominate the scene on Stack Overflow (~40%), followed by SQL (~20%), HTML (~18%), and Go-related (~12%) issues. In Reddit, there is no extreme gap across all technologies, but, in this case, the most frequent language is Go (~25%), still followed by SQL (~19%) and JavaScript (~15%). As before, questions about the remaining languages follow different patterns on the two Q&A websites. Overall, these results suggest that Stack Overflow is a referring platform for asking questions related to the most commonly used programming languages, such as Python and

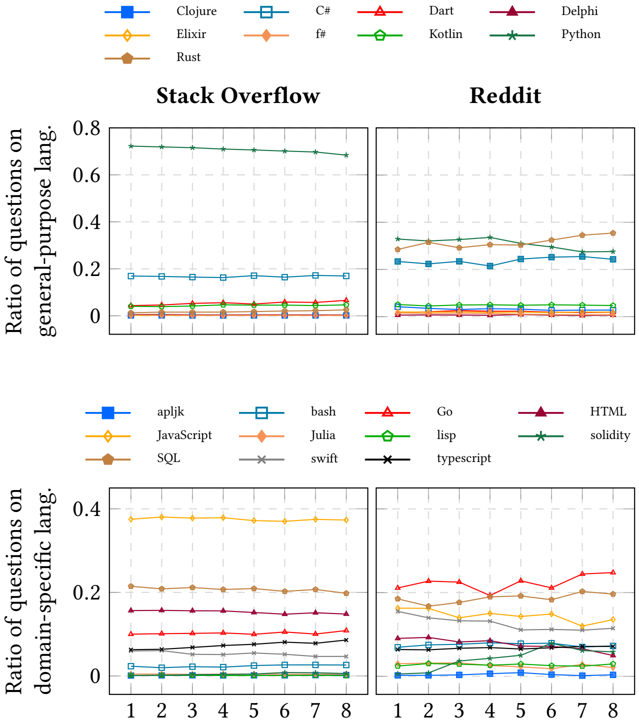


Figure 3: Proportion of the number of each language-related questions in Stack Overflow and Reddit.

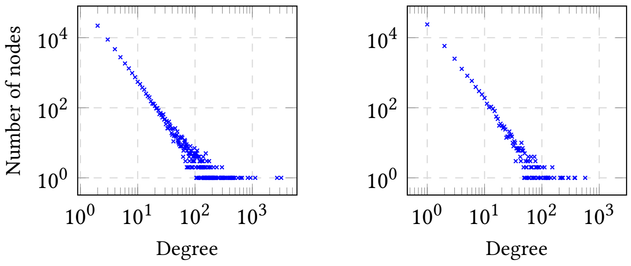
JavaScript. On the contrary, Reddit appears to be preferred to discuss recently rising programming languages and their applications in the software development landscape (e.g., Rust). Once again, this trend may be due to the inherent nature of the Reddit website, where users can either discuss or ask for help on a technical topic.

5.2 Mining users’ interaction patterns

Another distinctive aspect of how a Q&A platform is perceived by its users is how they spontaneously organize into informal communities. Understanding these dynamics would shed light not only on which technologies are more exploited for specific tasks but also on the platform’s durability over time and content quality.

5.2.1 Users-Questions interaction hypergraph. As described in Section 4, we built one hypergraph per trimester for each platform representing the many-to-many interaction between users commenting on the same questions. Figure 1 indirectly reports the size of each hypergraph, indicating the number of users (i.e., vertices) and questions (i.e., hyperedges) throughout the observation period. Figure 4 shows the vertex degree distribution of the Stack Overflow and Reddit hypergraphs built upon the first trimester of data. Recalling that evaluating the degree of a vertex in a so-built hypergraph translates to counting the number of discussions a user has contributed, it is not surprising that both Q&A websites exhibit a similar contribution pattern. This long-tail distribution is typical of online social platforms, where most of the users contribute poorly to the creation of new content, while only a few of them account for most of it, according to the online participation inequality pyramid [6, 24, 26]. In both platforms, we have that, on average, around the 65% of users have contributed to a single question/submission (69% in Stack Overflow, 64% Reddit), the 15% to two discussions (14% in Stack Overflow, 16% Reddit), the 6% to three questions (6% in Stack Overflow, 7% Reddit), and less than 12% to four or more discussions (11% in Stack Overflow, 13% Reddit). We found the same contribution pattern repeating over all trimesters in both platforms with no statistical difference.

5.2.2 Community detection and characterization. After building the user-question interaction hypergraphs, we evaluated their community structure via the label propagation algorithm (see Section 4). The upper part of Table 1 lists the number of communities found in each trimester for Stack Overflow (SO) and Reddit. The first remarkable outcome consists of the considerable number of communities discovered in both Q&A platforms, which decreases up to 20% for Stack Overflow while oscillating around the same value for Reddit. Further, the number of Reddit communities is always less than Stack Overflow communities by one order of magnitude throughout all trimesters. Although interesting, this result is not surprising and is consistent with previous literature. In a nutshell, the Reddit discussion model represents a fertile field for the growth of more supportive communities [2, 12, 17] whose interaction network appears to be less assortative and less clustered and, hence, less fragmented into small groups [15, 21]. On the contrary, Stack Overflow slowly changed over the last decade from a generalized discussion, with a small number of more extensive and diverse communities, to a more specialized mode with a larger number of smaller communities, each with a particular focus [23]. Having a deeper look at the size of the communities identified, we found that most of them were composed of very tiny groups of users and that, on average, only 5% of Stack Overflow and Reddit communities had a size bigger than 3 and 10, respectively. To filter out the noise such small groups could convey, we only focused on the top 5% largest communities on both platforms. The bottom part of Table 1 reports the percentage of users included in the communities considered: while we retained half of the original Reddit user base, we left with only 25% of the starting user base in the case of Stack Overflow.



(a) Stack Overflow.

(b) Reddit.

Figure 4: Vertex degree distribution, first trimester.

Number of communities across all trimesters								
SO	59179	58564	56181	52382	50838	49936	48979	47417
Reddit	5140	5890	5478	4703	4894	5414	5851	5311
% of users within the top 5% biggest communities								
SO	26,70%	26,05%	27,40%	26,29%	25,90%	25,61%	25,96%	26,44%
Reddit	52,07%	43,28%	50,37%	50,98%	52,79%	52,74%	47,53%	53,78%
Trim.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**

Table 1: Community and user distributions.

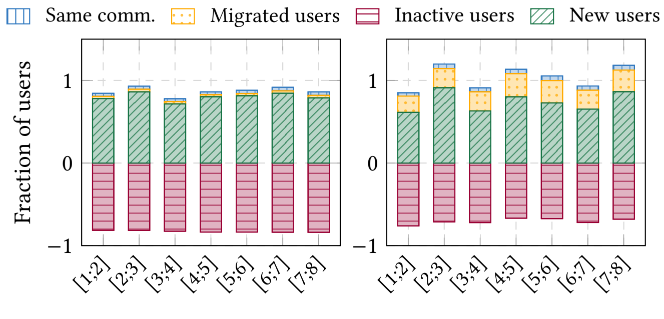
To characterize the most dominant topic within each community, we considered the most common tag among the hyperedges belonging to it. We then labeled each user with the same topic of the community they belong. Table 2 reports the aggregated percentage of users associated with a given programming language community in Stack Overflow and Reddit. Table 4 in Appendix A lists the size of the communities for all programming languages. These values can be interpreted as a complement to the trends of language-related questions shown in Figure 3. In Section 5.1.2, we focused on the number of questions per programming language; in this case, we consider the label associated with each user as a proxy

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

to count the number of users who have contributed to each topic. As expected, the biggest communities are represented by the most trending programming languages, such as Python and Javascript for Stack Overflow and Python, Rust, and Go for Reddit. Interestingly, we can observe how this pattern is not always true for Reddit. For instance, Clojure is the third most trending general-purpose language in our data set; however, less than 1% of the Reddit users analyzed belong to the Clojure community (a similar case happens for StackOverflow and C#). A completely reversed situation occurs for C#, where the C# community contains an average of 13% of the Reddit users. This result may suggest the existence of a few committed users who heavily discuss a given language (e.g., the Clojure community in Reddit) as well as the lack of solid interaction concerning a specific topic (e.g., C#-related questions in Reddit). Nevertheless, further analyses are needed to explain this behavior.

5.2.3 Community evolution. To verify how the composition of the communities changed over time, we examined 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 [6, 23]), or (re-)join the community (i.e., newly registered users or users that become active again after having been inactive the previous trimester(s)). Figure 5 shows the percentage of these four classes of users evaluated pairwise over consecutive trimesters. The values shown are averaged over all communities. What clearly stands out from this analysis is the significant number of users that become inactive from one trimester to the following (~82% for Stack Overflow, ~70% for Reddit) as well as the high number of new users (~80% for Stack Overflow, ~74% for Reddit). Further, it is also noticeable how the percentage of users remaining in the same community or migrating from one community to another is extremely low in the Stack Overflow platform (~4% and ~3%, respectively). At the same time, we can observe slightly higher values for Reddit (~5% and ~25%, respectively). Overall, these results suggest that most Stack Overflow users are interested in receiving feedback for solving their problems rather than contributing and building a community around specific topics. However, it is worth noting that this outcome is strictly related to the division in trimesters of the observation period and that this phenomenon may change if we observe a user's activity over more extended periods, for instance, from one year to another [23]. In the case of Reddit, we can observe a more significant portion of users (around 30% considering both users remaining in the same community and migrating to another) that consistently contribute to the platform. This behavior may reflect a greater sense of bonding among community members in addition to task-specific discussions [33]. Still, the migration pattern across communities must be further analyzed by considering each community independently.



(a) Stack Overflow. **(b) Reddit.**

Figure 5: Community evolution.

6 CONCLUSION

In this work, we have examined the developers' behavior and community formation around different programming languages in two popular Q&A platforms, namely Stack Overflow and Reddit. Our analysis highlighted a profound discrepancy in how these Q&A websites are utilized by their users as they differ in the length and duration of conversations, trending technologies, as well as community formation, and evolution. This distinction between Stack Overflow and Reddit results from their inherent design choices and social mechanics, which, in turn, influence their popularity and ability to create and maintain an active community of contributors. In line with previous literature comparing both platforms [20, 33, 40], our study emphasizes how Stack Overflow is falling out in favor of other Q&A websites, probably because of its gamification mechanism [29], the increasing amount of low-quality content [35], and the 'copy-and-paste' intent of its users, which are more focused on finding a working solution of their problems rather than contributing to new knowledge [9]. At the same time, we can observe the rise of Reddit as a comfort platform for addressing the more subjective and less technical discussion needs of the (lay) community [40], where users tend to remain engaged over time. Again, this pattern is probably motivated by Reddit's nature of being a discussion website, where the exchange of opinions is encouraged [17, 42]. Another point in favor of Reddit could also be that each subreddit chooses its internal rules with complete autonomy: in this way, each Reddit user can freely and transparently decide to which community they want to contribute.

The biggest potential threat to our work's validity resides in data gathering and processing. In the case of Stack Overflow, we filtered the data based on question tags explicitly naming the chosen programming languages. This process could have filtered out some questions that were not flagged as expected (e.g., a question about a cloud service tagged with the name of the cloud provider rather than the name of the language). In the case of Reddit, we could have included general submissions unrelated to the programming language (e.g., a moderator's post about rules in the subreddit). Further, we encountered some problems with the Pushshift API which could have undermined the exhaustiveness of our data set [8]. Nevertheless, we think that most of the programming language-related questions on those two sites are included in our study. Lastly, we did not filter out users based on their activity to not introduce selection bias towards the most active.

Future work will encompass two main directions. On one side, we will examine a more extensive user base spanning more than two years to grasp user behavior changes and community evolution considering different time granularities. On the other side, we aim to conduct a more detailed characterization of each community structurally (e.g., evaluating hypergraph motifs) 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 [13] and ChatGPT [27] to unravel how these trending tools are replacing standard Q&A websites in daily use.

REFERENCES

[1] A. Ahmad, C. Feng, S. Ge, and A. Yousif. 2018. A survey on mining stack overflow: question and answering (Q&A) community. *Data Technologies and Applications* 52, 2 (2018), 190–247. https://doi.org/10.1108/DTA-07-2017-0054

[2] M. Aniche, C. Treude, I. Steinmacher, I. Wiese, G. Pinto, M.-A. Storey, and M.A. Gerosa. 2018. How Modern News Aggregators Help Development Communities Shape and Share Knowledge. *Procs. - International Conference on Software Engineering* 2018-January, 499–510. https://doi.org/10.1145/3180155.3180180

[3] A. Antelmi, G. Cordasco, D. De Vinco, and C. Spagnuolo. 2023. Github Repository. https://github.com/ddevin96/DevCommunities. [Last accessed: March 2023].

[4] A. Antelmi, G. Cordasco, D. De Vinco, and C. Spagnuolo. 2023. Zenodo Dataset. https://zenodo.org/record/7685062. [Last accessed: March 2023].

[5] A. Antelmi, G. Cordasco, B. Kamiński, P. Prałat, V. Scarano, C. Spagnuolo, and P. Szufel. 2020. Analyzing, Exploring, and Visualizing Complex Networks via Hypergraphs using SimpleHypergraphs.jl. *Internet Mathematics* (2020). https://doi.org/10.24166/im.01.2020

[6] A. Antelmi, D. Malandrino, and V. Scarano. 2019. Characterizing the behavioral evolution of Twitter users and the truth behind the 90-9-1 rule. *The Web Conference 2019 - Companion of the World Wide Web Conference (WWW 2019)*, 1035–1038. https://doi.org/10.1145/3308560.3316705

[7] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri. 2020. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports* 874 (2020), 1–92. https://doi.org/10.1016/j.physrep.2020.05.004

[8] JSON Baumgartner. 2022. Pushshift API issues. https://twitter.com/jasonbaumgartne/status/1602767522868219905?cxt=HHwWgsC-ob-klr4sAAAA. [Last accessed: March 2023].

[9] G. Blanco, R. Pérez-López, F. Fdez-Riverola, and A.M.G. Lourenço. 2020. Understanding the social evolution of the Java community in Stack Overflow: A 10-year study of developer interactions. *Future Generation Computer Systems* 105 (2020), 446–454. https://doi.org/10.1016/j.future.2019.12.021

[10] A Bretto. 2013. *Hypergraph Theory: An Introduction*. Springer International Publishing, New York, NY (USA).

[11] C. Fu, X. Yue, B. Shen, S. Yu, and Y. Min. 2022. Patterns of interest change in stack overflow. *Scientific Reports* 12, 1 (2022). https://doi.org/10.1038/s41598-022-15724-3

[12] R.P. Gauthier, M.J. Costello, and J.R. Wallace. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3502076

[13] Github. 2022. Github Copilot. https://github.com/features/copilot. [Last accessed: March 2023].

[14] HackerNews. 2007. https://news.ycombinator.com/. [Last accessed: March 2023].

[15] W.L. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. 2017. Loyalty in online communities. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 540–543.

[16] Kaggle. 2022. Kaggle dataset for Stackoverflow platform. https://www.kaggle.com/datasets/stackoverflow/stackoverflow?datasetId=2268. [Last accessed: March 2023].

[17] Y. Liu and M. Anwar. 2022. Learning Programming in Social Media: An NLP-powered Reddit Study. *Proceedings - 2022 4th International Conference on Transdisciplinary AI, TransAI 2022*, 55–58. https://doi.org/10.1109/TransAI54797.2022.00015

[18] C. Lu, J. Peltonen, and T. Nummenmaa. 2019. Game postmortems vs. developer Reddit AMAs: Computational analysis of developer communication. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3337722.3337727

[19] Y. Lu, X. Mao, M. Zhou, Y. Zhang, T. Wang, and Z. Li. 2020. Haste Makes Waste: An Empirical Study of Fast Answers in Stack Overflow. *Proceedings - 2020 IEEE International Conference on Software Maintenance and Evolution, ICSME 2020*, 23–34. https://doi.org/10.1109/ICSME46990.2020.00013

[20] Y. Luo, P. Liang, C. Wang, M. Shahin, and J. Zhan. 2021. Characteristics and challenges of low-code development: The practitioners perspective. *International Symposium on Empirical Software Engineering and Measurement*. https://doi.org/10.1145/3475716.3475782

[21] A.N. Medvedev, R. Lambiotte, and J.-C. Delvenne. 2019. The Anatomy of Reddit: An Overview of Academic Research. *Springer Proceedings in Complexity*, 183–204. https://doi.org/10.1007/978-3-030-14683-2_9

[22] S. Meldrum, S.A. Licorish, and B.T.R. Savarimuthu. 2017. Crowdsourced knowledge on stack overflow: A systematic mapping study. *ACM International Conference Proceeding Series* Part F128635, 180–185. https://doi.org/10.1145/3084226.3084267

[23] I. Moutidis and H.T.P. Williams. 2021. Community evolution on stack overflow. *PLoS ONE* 16, 6 June 2021 (2021). https://doi.org/10.1371/journal.pone.0253010

[24] A. Nadiri and F.W. Takes. 2022. A Large-scale Temporal Analysis of User Lifespan Durability on the Reddit Social Media Platform. *WWW 2022 - Companion Proceedings of the Web Conference 2022*, 677–685. https://doi.org/10.1145/3487553.3524699

[25] L. Neuhäuser, A. Mellor, and R. Lambiotte. 2020. Multibody interactions and nonlinear consensus dynamics on networked systems. *Physical Review E* 101, 3 (2020). https://doi.org/10.1103/PhysRevE.101.032310

[26] J. Nielsen. 2006. The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities. https://www.nngroup.com/articles/participation-inequality/. Last accessed: March 2023.

[27] OpenAI. 2023. ChatGPT website. https://chat.openai.com/. [Last accessed: March 2023].

[28] Stack Overflow. 2008. https://stackoverflow.com/. [Last accessed: March 2023].

[29] M. Papoutsoglou, G.M. Kapitsaki, and L. Angelis. 2020. Modeling the effect of the badges gamification mechanism on personality traits of Stack Overflow users. *Simulation Modelling Practice and Theory* 105 (2020), 102157. https://doi.org/10.1016/j.simpat.2020.102157

[30] Pushshift. 2023. API to retrieve data from Reddit dump. https://api.pushshift.io. [Last accessed: March 2023].

[31] Reddit. 2005. https://reddit.com/. [Last accessed: March 2023].

[32] A. Saxena and H. Reddy. 2022. Users roles identification on online crowdsourced Q&A platforms and encyclopedias: a survey. *Journal of Computational Social Science* 5, 1 (2022), 285–317. https://doi.org/10.1007/s42001-021-00125-9

[33] S. Sengupta. 2020. 'Learning to code in a virtual world': A preliminary comparative analysis of discourse and learning in two online programming communities. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 389–394. https://doi.org/10.1145/3406865.3418319

[34] I. Srba and M. Bielikova. 2016. Why is Stack Overflow Failing? Preserving Sustainability in Community Question Answering. *IEEE Software* 33, 4 (2016), 80–89. https://doi.org/10.1109/MS.2016.34

[35] I. Srba and M. Bielikova. 2016. Why is Stack Overflow Failing? Preserving Sustainability in Community Question Answering. *IEEE Software* 33, 4 (2016), 80–89. https://doi.org/10.1109/MS.2016.34

[36] Stackoverflow. 2022. annual survey of Stackoverflow platform. https://survey.stackoverflow.co/2022. [Last accessed: March 2023].

[37] StartupBonsai. 2023. 23+ Reddit Statistics For 2023: Users, Revenue, And Growth. https://startupbonsai.com/reddit-statistics. [Last accessed: March 2023].

[38] J. Surowiecki. 2016. The Wisdom of Crowds. 37 (2016), 351–354. Issue 3.

[39] I. Waller and A. Anderson. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 1954–1964. https://doi.org/10.1145/3308558.3313729

[40] M. Wu, R. Aranovich, and V. Filkov. 2021. Evolution and differentiation of the cybersecurity communities in three social question and answer sites: A mixed-methods analysis. *PLoS ONE* 16, 12 December (2021). https://doi.org/10.1371/journal.pone.0261954

[41] D. Yang, B. Qu, J. Yang, and P. Cudre-Mauroux. 2019. Revisiting User Mobility and Social Relationships in LBSNs: A Hypergraph Embedding Approach. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, 2147–2157. https://doi.org/10.1145/3308558.3313635

[42] A. Zeng and W. Crichton. 2019. Identifying barriers to adoption for rust through online discourse. *OpenAccess Series in Informatics* 67. https://doi.org/10.4230/OASIcs.PLATEAU.2018.5

A APPENDIX 1

Table 3 shows the distribution of the size of Stack Overflow and Reddit communities. Table 4 reports the aggregated percentage of users associated with a given programming language community in Stack Overflow and Reddit. This table extends Table 2 by listing the data associated with all programming languages.

		Percentiles			
Trimester	**Q&A Site**	**0.25**	**0.5**	**0.75**	**0.95**
1	Stack Overflow	1	1	2	3
	Reddit	1	2	5	11
2	Stack Overflow	1	1	2	3
	Reddit	1	2	5	12
3	Stack Overflow	1	1	2	3
	Reddit	1	2	5	12
4	Stack Overflow	1	1	2	3
	Reddit	1	2	5	12
5	Stack Overflow	1	1	2	3
	Reddit	1	2	5	12
6	Stack Overflow	1	1	2	3
	Reddit	1	2	5	11
7	Stack Overflow	1	1	2	3
	Reddit	1	2	5	13
8	Stack Overflow	1	1	2	3
	Reddit	1	2	5	13

Table 3: Distribution of the size of Stack Overflow and Reddit communities.

		Trimester							
Prog. lang.	**Q&A Site**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
APL	Stack Overflow	0,00	0,03	0,08	0,00	0,00	0,00	0,06	0,02
	Reddit	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Bash	Stack Overflow	0,23	0,44	0,30	0,46	0,51	0,71	0,73	0,74
	Reddit	1,25	1,06	2,11	1,38	0,47	1,88	1,60	1,05
C#	Stack Overflow	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	Stack Overflow	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
Dart	Stack Overflow	0,77	1,06	1,05	0,57	0,77	1,21	1,16	1,16
	Reddit	0,15	0,49	0,57	0,14	0,13	0,32	0,27	0,32
Delphi	Stack Overflow	0,14	0,06	0,09	0,11	0,02	0,11	0,02	0,02
	Reddit	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,00
Elixir	Stack Overflow	0,06	0,01	0,02	0,00	0,01	0,02	0,00	0,02
	Reddit	0,67	0,57	0,67	0,65	0,18	0,21	0,88	0,47
f#	Stack Overflow	0,05	0,03	0,05	0,04	0,02	0,04	0,01	0,00
	Reddit	0,12	0,00	0,00	0,00	0,00	0,00	0,00	0,19
Go	Stack Overflow	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
HTML	Stack Overflow	1,55	1,72	1,36	1,55	1,57	1,59	1,65	1,05
	Reddit	0,11	0,00	0,45	0,38	0,29	0,32	0,55	0,00
Javascript	Stack Overflow	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
Julia	Stack Overflow	0,06	0,29	0,16	0,15	0,12	0,04	0,04	0,09
	Reddit	0,40	0,92	2,47	0,44	0,35	0,42	0,40	0,29
Kotlin	Stack Overflow	0,70	0,99	0,45	1,11	0,80	0,83	0,82	0,90
	Reddit	1,71	1,53	1,17	0,42	0,93	1,77	1,13	1,06
Lisp	Stack Overflow	0,07	0,07	0,00	0,04	0,00	0,02	0,08	0,00
	Reddit	1,41	0,67	1,03	1,58	1,50	0,71	0,51	0,82
Python	Stack Overflow	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	Stack Overflow	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
Solidity	Stack Overflow	0,00	0,02	0,03	0,16	0,15	0,15	0,10	0,00
	Reddit	0,00	0,00	0,11	0,19	0,14	0,99	0,00	0,59
SQL	Stack Overflow	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
Swift	Stack Overflow	1,92	1,71	1,13	0,92	1,01	1,58	0,78	1,04
	Reddit	2,44	3,62	2,28	1,78	1,78	1,81	1,24	1,87
Typescript	Stack Overflow	0,55	0,62	0,42	0,93	0,54	0,70	0,62	0,75
	Reddit	1,40	1,15	1,12	1,47	0,85	1,21	1,19	1,98

Table 4: Percentage of users per programming language within the top 5% biggest communities.