Reading Answers on Stack Overflow: Not Enough!

Haoxiang Zhang, *Member, IEEE*, Shaowei Wang, *Member, IEEE*, Tse-Hsun (Peter) Chen, and Ahmed E. Hassan, *Fellow, IEEE*

Abstract—Stack Overflow is one of the most active communities for developers to share their programming knowledge. Answers posted on Stack Overflow help developers solve issues during software development. In addition to posting answers, users can also post comments to further discuss their associated answers. As of Aug 2017, there are 32.3 million comments that are associated with answers, forming a large collection of crowdsourced repository of knowledge on top of the commonly-studied Stack Overflow answers. In this study, we wish to understand how the commenting activities contribute to the crowdsourced knowledge. We investigate what users discuss in comments, and analyze the characteristics of the commenting dynamics, (i.e., the timing of commenting activities and the roles of commenters). We find that: 1) the majority of comments are informative and thus can enhance their associated answers from a diverse range of perspectives. However, some comments contain content that is discouraged by Stack Overflow. 2) The majority of commenting activities occur after the acceptance of an answer. More than half of the comments are fast responses occurring within one day of the creation of an answer, while later comments tend to be more informative. Most comments are rarely integrated back into their associated answers, even though such comments are informative. 3) Insiders (i.e., users who posted questions/answers before posting a comment in a question thread) post the majority of comments within one month, and outsiders (i.e., users who never posted any question/answer before posting a comment) post the majority of comments after one month. Inexperienced users tend to raise limitations and concerns while experienced users tend to enhance the answer through commenting. Our study provides insights into the commenting activities in terms of their content, timing, and the individuals who perform the commenting. For the purpose of long-term knowledge maintenance and effective information retrieval for developers, we also provide actionable suggestions to encourage Stack Overflow users/engineers/moderators to leverage our insights for enhancing the current Stack Overflow commenting system for improving the maintenance and organization of the crowdsourced knowledge.

Index Terms—Crowdsourced Knowledge Sharing and Management, Stack Overflow, Commenting, Empirical Software Engineering

1 Introduction

Stack Overflow has changed how developers ask and answer programming-related questions. Stack Overflow provides a knowledge sharing platform to help developers share knowledge and seek answers to their problems. Stack Overflow has accumulated a large amount of programming knowledge in the form of questions and answers. As of Aug 2017, there are 14.5 million questions on Stack Overflow. The question answering activities cover various software development domains, and have generated 22.7 million answers. Stack Overflow has become a community with 7.6 million registered users who contribute and share their programming knowledge in a crowdsourced manner.

Such crowdsourced knowledge is not only generated by the question answering process, but it is also produced by commenting activities. Comments are appended under their associated questions and answers to facilitate further discussion[1]. For example, in Fig. 1, a comment is posted to provide additional information (i.e., the limitation of *overflow:hidden*) to its associated answer.

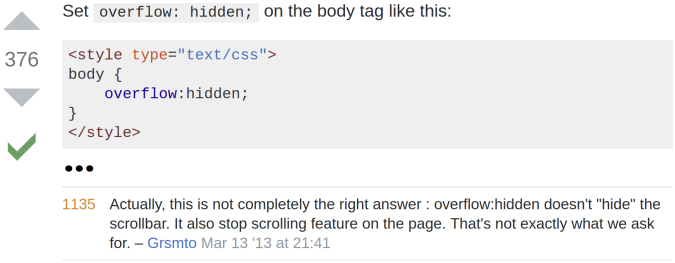

Fig. 1: An example of a comment that is associated with an answer. This comment points out a flaw in the accepted answer and has gained a higher score (i.e., 1,135) than its associated answer (i.e., 376).

Although commenting is a popular online communication channel, commenting activities on Stack Overflow have never been studied in depth before. Note that posting comments on Stack Overflow does not generate any reputation point for a user. From a user's perspective, comments can be easily missed in contrast to answers. Hence, Stack

1. Note that we refer to comments that are associated with answers on Stack Overflow as comments, if not specified otherwise in the rest of the paper. Comments that are associated with questions are not studied in this paper.

- *H. Zhang and A. E. Hassan are with the Software Analysis and Intelligence Lab (SAIL), Queen's University, Kingston, Ontario, Canada. E-mail: hzhang,ahmed@cs.queensu.ca*
- *S. Wang is with the Department of Computer Science and Engineering, Mississippi State University, Mississippi State, Mississippi, United States. E-mail: wang@cse.msstate.edu*
- *T. Chen is with the Software PErformance, Analysis, and Reliability (SPEAR) lab, Concordia University, Montreal, Quebec, Canada. E-mail: peterc@encs.concordia.ca*
- *Shaowei Wang is the corresponding author.*

Overflow has suggestions about what should (e.g., request clarification and leave constructive criticism) and should not (e.g., answer a question and suggest corrections) be posted in comments. On the other hand, users may not know Stack Overflow's commenting guideline. For example, on Stack Overflow META[2], some users hold an opinion[3] that knowledge sharing should only be conducted in the form of answers and not comments (in Fig. 2). Others consider comments as *temporary "Post-It" notes* to improve their associated answers[4], as shown in Fig. 3. A prior study shows that certain comments contain informative content. For instance, comments can point out the obsolescence of their associated answers [1]. However, no systematic study has ever been done on the comments of Stack Overflow to better understand how comments are used. For instance, whether users use comments by following Stack Overflow commenting guidelines?

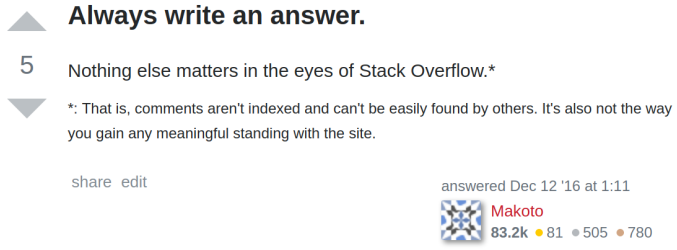

Fig. 2: A discussion on Stack Overflow META shows an opinion that knowledge sharing should exclusively occur in answers.

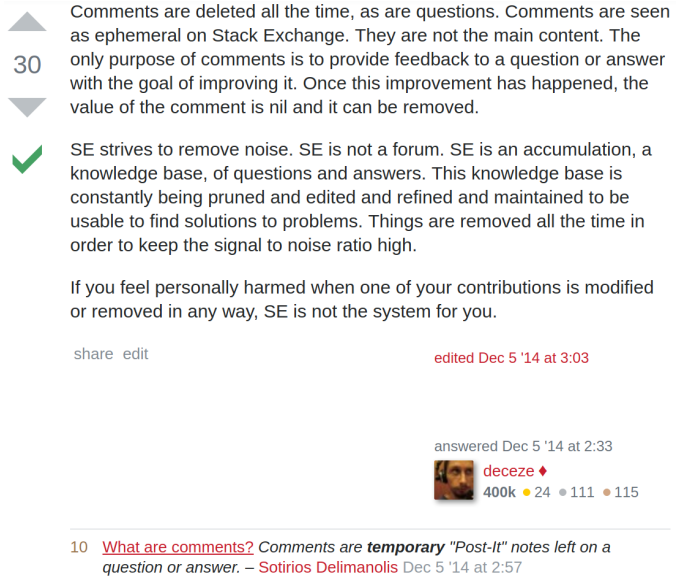

Fig. 3: A discussion on Stack Overflow META shows an opinion that comments are temporary.

Therefore, in this paper, we investigate the comments (i.e., 32.3 million) that are associated with Stack Overflow answers (i.e., 22.7 million) to gain a better understanding of how the commenting activities contribute to the generation and maintenance of crowdsourced knowledge. We wish to provide insights to Stack Overflow users so that they can more effectively identify relevant information from comments. We also wish to provide actionable suggestions to Stack Overflow designers and engineers so that the Stack Overflow commenting system can be used more effectively to enhance the knowledge sharing process.

More specifically, we first conduct a preliminary study on how active users are in posting comments. We find that a large collection of comments exist, and that the number of posted comments continues to exceed the number of posted answers every year since 2009. 23% (i.e., 2.6 million) of the answers with comments have a commenting-thread (i.e., all the comments that are associated with an answer) that is longer than the actual answer. Then, we answer the following three RQs:

- **RQ1: What do users discuss in comments?**

Most comments (i.e., 75.6%) provide useful information, such as pointing out weaknesses and providing references to their associated answer. However, some of the informative comments (e.g., suggesting a correction and answering a question) do not follow the commenting guidelines that are outlined by Stack Overflow.

- **RQ2: When do users post comments?**

The acceptance of an answer is not the end of commenting activities; instead, the majority of commenting activities occur after the acceptance of an answer. Generally, more than half of the comments are fast responses occurring within one day of the answer creation. Comments that point out the advantage and weakness of answers tend to be posted later than other informative comments, and later comments tend to be more informative. However, the knowledge in comments is rarely integrated back into answers.

- **RQ3: What types of users participate in the commenting activities?**

Users are highly involved in commenting. Askers mainly comment to express praise, inquiry, and point out weakness, while answerers mainly comment to highlight the advantage of an answer, provide improvement and additional information. Askers and answerers are more likely to post comments within one month, while other users are more likely to post comments after one month. Among informative comments, inexperienced users tend to raise limitations and concerns while experienced users tend to enhance answers by commenting. Among uninformative comments, inexperienced users tend to praise answers while experienced users tend to post irrelevant information.

Based on our findings, we encourage users to read comments carefully since the majority of comments provide a diverse variety of information that enhances their associated answers. Thus, comments are an important resource of knowledge. For example, a comment can provide helpful clarification to its associated answer, or point out flaws. Especially, we highlight that later comments tend to be more informative than comments that are posted soon after the posting of their associated answer. However, the informative content in comments is rarely integrated back into their associated answers. Thus,

2. Stack Overflow META is the part of the site where users discuss the workings and policies of Stack Overflow.
3. https://meta.stackoverflow.com/a/339395
4. https://meta.stackoverflow.com/a/278517

Stack Overflow should consider adopting a mechanism to reward reputation points or certain badges to encourage the maintenance and integration of commenting knowledge. We also suggest that Stack Overflow designers should improve the current commenting system because users post comments in unrecommended manners (i.e., not following Stack Overflow's commenting guideline), such as suggesting corrections, answering questions, and praising answers.

Paper Organization: The rest of this paper is organized as follows. Section 2 introduces the background of the commenting system on Stack Overflow. Section 3 describes our studied dataset, and explores the commenting activities on Stack Overflow as a preliminary study. Section 4 details the results from our case study. Section 5 discusses our findings and their implications. Section 6 discusses the potential threats to the validity of our findings. Section 7 surveys relevant work to our study. Finally, Section 8 concludes our study.

2 Background

2.1 Question answering on Stack Overflow

Stack Overflow is an online platform for question answering in the domain of software programming. Users can post their questions with descriptive text about their problems. Once a question is posted by a user (i.e., an asker), others (i.e., answerers) can post answers to this question. When an asker is satisfied with any solution, she/he can select the answer as the accepted answer among all the posted answers.

A reputation system is implemented on Stack Overflow to measure and encourage the contributions of users to the community. There are various ways for users to gain reputation points. For instance, an answer can be upvoted by other users; as a result, the answerer gains 10 reputation points as reward. On the other hand, commenting activities do not lead to any gain of reputation points, since they "*are all secondary activities*" according to Stack Overflow[5].

2.2 Commenting on Stack Overflow

Users (i.e., commenters) can post additional text/code under either questions or answers. As stated by Stack Overflow, "*comments are temporary 'Post-It' notes left on a question or answer ... but do not generate reputation*" [2]. We consider a question with all its answers and their associated comments as a *question thread*. Similarly, we consider an answer with its associated comments as an *answer thread*, and all the comments that are associated with an answer as a *commenting-thread*.

Comments can be posted by three types of users [2]:

- The owner of an answer can post comments under the answer;
- The owner of a question can post comments under the question and any of its answers;
- Users with at least 50 reputation points can post comments everywhere.

5. https://stackoverflow.com/help/whats-reputation

Similar to answers, comments can be also upvoted; however, upvoting a comment does not generate any reputation points. Officially, Stack Overflow recommends users to post comments under the following circumstances: ***request clarification***, ***leave constructive criticism***, and ***add relevant information***, and recommends users *not to* post comments under circumstances such as: ***suggest corrections***, ***answer a question***, and ***compliment*** [2]. In the following RQs, we investigate the content of comments to examine if users follow Stack Overflow commenting guidelines.

3 Data Collection and Preliminary Study

Stack Overflow is the largest Q&A website tailored for software developers, with 7.6 million registered users. As of Aug 2017, 14.5 million questions have been asked across more than 5,000 tags (i.e., user-provided topics of a question). Developers leverage answers posted on Stack Overflow to tackle their coding issues or learn programming knowledge. Similar to other online platforms (e.g., Reddit, Hacker News, and Quora), users can also post comments to answers. In order to get an understanding of the commenting activities on Stack Overflow, we first conduct a preliminary study on such comments. More specifically, we wish to find the popularity of commenting activities, considering that posting comments does not result in extra reputation points.

We downloaded all the comments from the Stack Overflow data dump on archive.org[6] that was released in Aug 2017. In this dataset, there are 60.1 million comments from either questions or answers. Since we focus on the comments that are associated with answers, we end up with 32.2 million comments in this study, which are associated with 22.7 million answers and 1.9 million users. In general, comments are short. The median length of a comment is 115 characters.

We compare the number of comments and answers posted on a yearly basis. We also characterize the amount of information in comments that are associated with answers. Since each comment is associated with an answer, we measure the number of characters in all the comments that are associated with an answer (i.e., a commenting-thread) and compare it with that of its associated answer.

Stack Overflow has a large collection of comments. More comments were posted than answers yearly since 2009. As of August 2017, among 11.4 million answers, 32.3 million comments are posted. The number of comments is greater than the number of answers on a yearly basis since 2009, as shown in Fig. 4. The ratio between the number of comments and answers keeps increasing until 2013, and remains stable afterwards[7]. The ratio between the number of comments and answers is around 1.5:1 since 2013. In other words, users are actively commenting on answers.

23% (i.e., 2.6 million) of the answers with comments have a commenting-thread that is longer than the actual answer (in terms of characters). In such answer threads,

6. https://archive.org/details/stackexchange

7. Note that we used the data dump that is published in Aug 2017; therefore, the numbers of both answers and comments in 2017 do not cover the entire year.

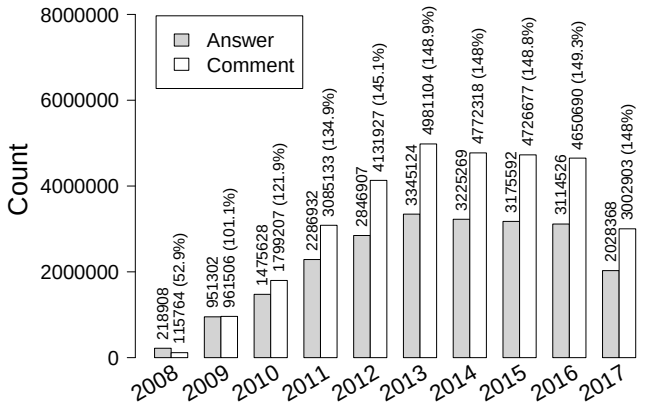

Fig. 4: The number of answers and comments created on a yearly basis. The percentage value shown in the bracket is the ratio of the number of comments to the number of answers on a yearly basis.

users may require more time and effort to read the comments than the answer, not only due to the longer text, but also due to the free style of comments and the way comments are organized and presented. A commenting-thread can be massive and lack organization, thus leading to information overload and hindering information retrieval. In an example[8], the answer has 3,175 characters, while the answer has 28 comments which have 5,460 characters in total. It is difficult to understand the conversations in the commenting-thread from the default Stack Overflow view of comments. Additionally, 23 comments are hidden from this view. Thus, it is ineffective to share and retrieve information in a commenting-thread, even when a comment can enhance its associated answer. Only from the view that displays all the comments, then it is more clear what is the context of an individual comment, and whom a commenter mentions in his/her conversation. The unorganized nature of comments increases the difficulty to read and understand conversations. Due to the above-mentioned obstacles, comprehending all the information requires extra time and effort. It can be intimidating that some users may not bother to read comments at all. This polarizing view of comments has been investigated by Reich [3]. In his study, one interviewee said that "*Most people don't want to comment. And actually, most people don't want to read other people's comments*". Therefore, we conduct an empirical study to gain a better understanding of what users discuss in comments, when users post comments, and who participate in such commenting activities.

Based on the above-mentioned preliminary results, we find that users actively participate in commenting activities, and sometimes comments that are associated with an answer can be longer than their associated answer – making it time-consuming to read comments. Therefore, in this paper, we wish to explore this large collection of comments. In the following RQs, we study how comments provide useful information to their associated answers, and understand the characteristics of commenting. We hope to provide insights on how to leverage these comments as well as to suggest mechanisms to more effectively organize comments for easy information retrieval and knowledge management.

> Stack Overflow has a large collection of comments, whose number is even larger than answers. The amount of information in comments cannot be neglected, with 23% of the answers having a commenting-thread that is longer than their actual answer.

4 Case Study

4.1 RQ1: What do users discuss in comments?

Motivation: Commenting activities on Stack Overflow open up an alternative channel for users to participate in the knowledge crowdsourcing process. As shown in Section 3, in 23% of answers with comments, the commenting-thread is even longer than the answer itself. Some comments significantly add value to their associated answers. For example, in a comment[9] shown in Fig. 1, the commenter pointed out that the accepted answer is not completely correct. Although the answer has been accepted by the asker and upvoted by the community to reach a score of 376, this comment has been upvoted even more and has a higher score (i.e., 1,135) than the associated answer. On the other hand, users can post comments in a relatively free style, and a commenting-thread can even appear unorganized.

Such a large collection of commentary text that is associated with answers is important for users but not well understood nor studied carefully. It is unclear what do users actually discuss in comments and whether or not Stack Overflow commenting guidelines (see Section 2) are followed by users. Therefore, in this RQ, we investigate what users actually discuss in comments. Moreover, we categorize the types of discussions in comments (i.e., the comment types), and investigate the advantages and disadvantages of different comment types with regard to the official guidelines from Stack Overflow. A better understanding of the comment types can provide Stack Overflow engineers with firsthand insights into how commenting as a communication channel is used in practice. The identified comment types can also be leveraged to better organize comments and improve the maintenance of crowdsourced knowledge so that answer seekers can effectively extract relevant information.

Approach: We study what do users discuss in comments through qualitative analysis. First, we randomly select a statistically representative sample of 3,000 comments from the 32.2 million comments, providing us with a confidence level of 99% and a confidence interval of 2.4%. We manually label the types of discussions in each comment at the sentence level with a lightweight open coding-like process. For example, in a comment[10], the user says that "*This one works and should be answer to this question. Although dot, coma and other values are displayed to user, the user can not insert them. So only input you receive from this is numeric.*". This comment is assigned the type *praise* and *clarification*. This lightweight open coding-like process involves 3 phases and

8. https://stackoverflow.com/a/47990

9. https://www.stackoverflow.com/posts/comments/21766075
10. https://www.stackoverflow.com/posts/comments/12535822

is performed by the first two authors (i.e., A1-A2) of this paper as follows:

- **Phase I:** A1 derives a draft list of comment types based on 50 randomly sampled comments. Then, A1 and A2 use the draft list to label the sampled comments collaboratively. During this phase the comment types are revised and refined.
- **Phase II:** A1 and A2 independently apply the resulting types from Phase I to label the rest of the 3,000 comments. A1 and A2 take notes regarding the deficiency or ambiguity of the labeling for the comments. Note that new labels (i.e., the comment types) are introduced during this phase if A1 or A2 observes more comment types. At the end of this phase, we obtain 7 types of comments that are further divided into 17 subtypes (see Table 1).
- **Phase III:** A1 and A2 discuss the coding results that are obtained in Phase II to resolve any disagreements until a consensus is reached. No new types and subtypes are added during this discussion. The inter-rater agreement of this coding process at the subtype and type level has a Cohens kappa of 0.86 and 0.90 (measured at the start of Phase III) respectively. These kappa levels indicate that the agreement level is substantial.

We analyze each comment type and present concrete examples. Furthermore, to evaluate how the actual commenting activities are aligned with the official commenting guidelines from Stack Overflow, we compare the recommended commenting scenarios with the actual commenting scenarios, and examine the reasons for agreements and disagreements.

Results: 75.6% of the tagged comments are informative. Table 1 shows the identified comment types with their subtypes. Users often post informative comments, i.e., comments of type *advantage*, *improvement*, *weakness*, *inquiry*, and *addition*, to enhance their associated answers. Users also post uninformative comments, i.e., comments of type *praise* and *irrelevant*. **Furthermore, diverse subtypes of comments exist for each informative comment type.** More specifically, 37.7% of the studied comments belong to type *addition*. These comments add value to their associated answers by providing new content, i.e., an alternative solution, a concrete example, clarification, or a reference link. For example, in Fig. 5, the comment[11] points out an update in TensorFlow and provides a reference link. 19.7% of the comments belong to the type *inquiry*. These comments pose additional questions that are related to their associated answers, or request extra information for better understanding their associated answer, such as asking the answerer where "JsonConvert" is originated in the code snippet of the answer[12]. Comments of type *inquiry* aim to motivate the answerers to disclose more details, it helps answers to become clearer, and thus, are more likely to be used by other users. 17.3% of the comments are of type *weakness*. In these comments, a weakness in an answer (i.e., flaws, coding errors, obsolescence, and disagreements) is noted. For example, in a question about issues converting a javascript object to a query string, a comment[13] points out that an answer "*... only do convert plain js object to query string. If you need to resolve for nested objects go with some recursive strategy*". Last but not least, 5.9% and 4.5% of the comments improve existing answers, and comment on the advantage of an existing answer, respectively. The uninformative comments either praise an answer (i.e., 17.4%, such as "*Thank you. It worked for me :)*"[14]) or discuss irrelevant topics (i.e., 14.7%, such as "*If it correct please vote up Thanks :)*"[15]). They are not informative because they do not directly add value to their associated answers.

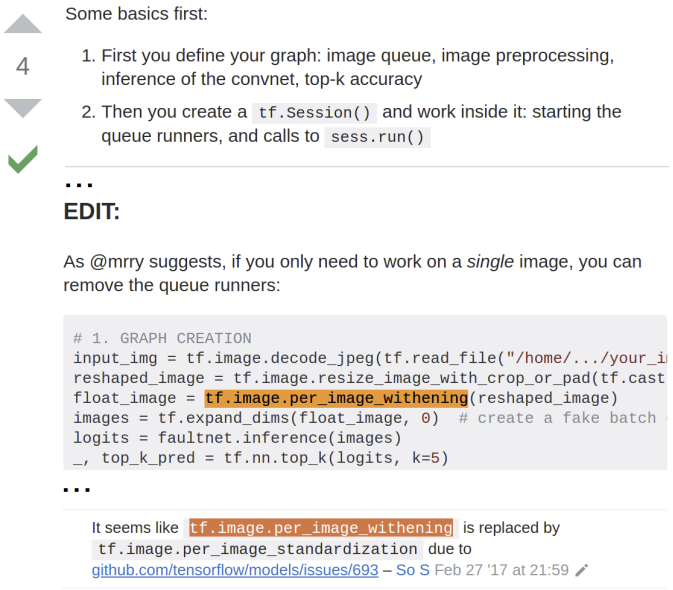

Fig. 5: An example of a comment that points out a TensorFlow update with a reference link.

The majority of informative comments (i.e., 67.4% of the tagged comments) follow the Stack Overflow commenting guidelines. Stack Overflow recommends users to post comments when they want to ***request clarifications*** from the author of posts. We observe that comments of type *inquiry* match this guideline. Comments of type *weakness* point out flaws, coding errors, obsolescence, or disagreements, thus also follow the guideline to ***leave constructive criticism***. Comments of type *addition* are encouraged by Stack Overflow as well because they ***add relevant information*** to an answer.

Comments that *suggest corrections*, *answer a question*, or are *compliment* exist (i.e., 31.3% of the tagged comments), although they are discouraged by Stack Overflow. For example, Stack Overflow does not recommend comments that ***suggest corrections***; however, comments of type *improvement* point out and take actions to the weakness in their associated answers, such as making corrections, or providing extensions or solutions, to fix an answer's weakness. Nevertheless, such comments are not recommended by Stack Overflow. Instead, users are recommended to suggest or make an edit to an existing answer. As shown in Fig. 12, the median reputation points for all comment types are below 2,000 (Stack Overflow only allows users with more than

11. https://www.stackoverflow.com/posts/comments/72132287
12. https://www.stackoverflow.com/posts/comments/65851467
13. https://www.stackoverflow.com/posts/comments/72503806
14. https://www.stackoverflow.com/posts/comments/74500847
15. https://www.stackoverflow.com/posts/comments/66524665

TABLE 1: The definition of types of comments with their subtypes and the proportion of each type.

Type	Count	Subtype	Count	Description
Praise	521 (17.4%)	Praise	521 (17.4%)	Praises the answer
Advantage	134 (4.5%)	Support	43 (1.4%)	Gives reasons to support the answer
		Highlight	50 (1.7%)	Highlights the working circumstances or use case of the answer
		Performance	43 (1.4%)	Discusses the performance of the answer
Improvement	176 (5.9%)	Correction	152 (5.1%)	Provides correction to the answer
		Extension	24 (0.8%)	Extends the answer to other cases
Weakness	518 (17.3%)	Flaw	238 (7.9%)	Points out flaws or limitations
		Error	139 (4.6%)	Points out errors in the code
		Obsolete	31 (1.0%)	Points out obsoleteness
		Disagree	112 (3.7%)	Disagrees with the answer
Inquiry	591 (19.7%)	Question	446 (14.9%)	Asks clarification questions
		Request	147 (4.9%)	Requests information
Addition	1,130 (37.7%)	Solution	293 (9.8%)	Provides alternative solutions to the answer
		Example	66 (2.2%)	Adds a concrete example
		Clarification	682 (22.7%)	Adds a clarification
		Reference	163 (5.4%)	Adds a reference
Irrelevant	441 (14.7%)	Irrelevant	441 (14.7%)	Discusses irrelevant topics to the answer

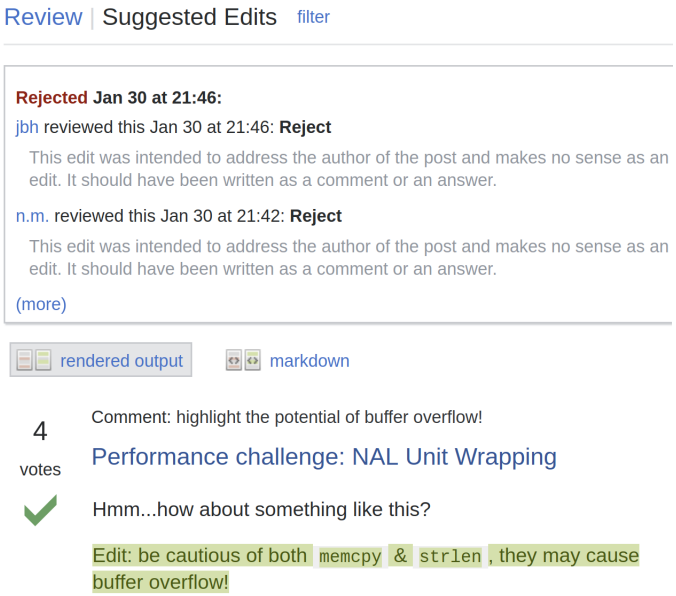

Fig. 6: A suggest edit with rejecting reasons.

2,000 reputation points to edit an answer directly). Note that when a user does not have enough reputation points (i.e., 2,000) to directly edit the answer, he/she can suggest an edit. Under this circumstance, the suggested edit is placed in a review queue[16]. However, it is unknown whether or not the suggested edit will be accepted, therefore the commenter faces uncertainty if he/she attempts to enhance an answer by editing directly. An example of a suggested edit[17] is shown in Fig. 6. The edit was rejected by two reviewers because: "*This edit was intended to address the author of the post and makes no sense as an edit. It should have been written as a comment or an answer*". The reviewers suggested the editor to write a comment or an answer to correct the answer instead of editing the answer. However, Stack Overflow does not recommend users to post comments to "*suggest corrections that don't fundamentally change the meaning of the post; instead, make or suggest an edit*". Additionally, it is unnecessary to create a new answer that simply corrects an existing answer. As a result, a contradictory situation occurs for such users with lower than 2,000 reputation points. Namely, ***edits that improve an answer can be rejected and suggested to become a comment, although, simultaneously, Stack Overflow does not recommend the posting of comments to suggest corrections – thus, users who wish to correct an answer are suggested not to edit the answer nor to post a comment, but to remain silent!*** As a matter of fact, posting a comment has a lower barrier, i.e., anyone with at least 50 reputation points can post comments, while any user with less than 2,000 reputation points has to wait for approval for editing an answer directly. Therefore, even while being discouraged to maintain an answer in the above scenario, users who still attempt to maintain the answer are likely to still post comments to suggest corrections. Even if a suggested edit is accepted, or the user has at least 2,000 reputation points, the answerer may not prefer others to change the post and might simply rollback the edit [4]. Overall, the above-mentioned obstacles create a cumbersome situation for users who wish to suggest corrections to an answer.

Such conflict also applies to the Stack Overflow commenting guideline that users should not post comments to ***answer a question***. Even though it is not recommended to answer a question in a comment, we still observe cases of posting an answer in a comment instead of editing an answer or creating a new answer. Another type of comments that is not recommended by Stack Overflow is comments of type *praise*, although users still praise in comments. In conclusion, we suggest the implementation of new mechanisms to tackle these issues. For example, answerers can be notified when comments are posted to correct their answers or answer new questions. Once commenters post information to enhance answers, the answerers or the community can decide whether or not to accept their effort for knowledge maintenance. Comments of type *praise* can also be detected automatically, and Stack Overflow can suggest these commenters to upvote the answer instead. Meanwhile, users can be provided with an option to post a short comment when they upvote an answer.

16. https://stackoverflow.com/help/editing
17. https://stackoverflow.com/review/suggested-edits/22075882

The majority of comments enhance their associated answer from a diverse range of perspectives (e.g., pointing out weaknesses or providing additional references). Even though the majority of informative comments follow Stack Overflow commenting guidelines (e.g., requesting clarification and adding relevant information), users are still posting a considerable portion of comments that are discouraged by Stack Overflow (e.g., praising an answer, or suggesting a correction).

4.2 RQ2: When do users post comments?

Motivation: Once an answer is created, comments can be posted under that answer at any time. Meanwhile, the answer can be edited to reflect any update to its content. In Section 4.1, we observed that the majority of comments are informative, and thus, can be potentially leveraged to enhance their associated answers. Therefore, in this RQ, we analyze the temporal dynamics of comments to find out when the commenting activities occur and when their associated answers are edited. Our findings may be an indicator of the ability of the community to effectively integrate comments back into their associated answers, because "*comments are temporary 'Post-It' notes left on a question or answer*" [2]. Ideally, the value of a comment can be best reflected in its associated answer. In addition, accepted answers are considered as the "best" solutions given to their associated questions [5]. Therefore, we wish to analyze how the timing of commenting activities is associated with the timing of answer acceptance. Understanding the temporal dynamics of comments provides us with insights into how to effectively manage the large collection of commenting-threads.

Approach: We first investigate when do users post comments. Namely, how long it takes for a user to post a comment since the creation of an answer, and whether a comment is posted before or after an answer is accepted. We also classify the commenting time into three categories: within one day, from one day to 30 days, and more than 30 days, to characterize whether or not a commenting activity occurs as a fast response to an answer. Among the accepted answers, we analyze whether the comment was posted before or after its associated answer was accepted. We analyze when do users post comments across different comment types that we identified from the 3,000 statistically representative samples of comments in Section 4.1 to understand the relationship between the type of comments and the timing of commenting. Lastly, from all the answers with comments, we extract the creation time and the last edit time of the answers. We also extract the creation time of the latest comments in these answers. By examining the proportion of answers that are edited after the discussion through comments, we wish to estimate an upper bound on the time it takes for the comments to be integrated back into their associated answers.

Results: Most comments were posted within one day after the creation of their associated answers. Comments of type *advantage* and *weakness* are more likely to be posted later than other informative comment types. The proportion of comments posted during different ranges of commenting time is shown in Fig. 7 for each comment type. More than half of the tagged comments are posted within one day for each comment type. Comments of type *irrelevant* and *improvement* have a higher chance to be posted within one day. Within one day that the answer is posted, 70.8% of the posted comments are informative. From day one to day three, 75.5% of the posted comments are informative. After three days, 78.4% of the posted comments are informative. After one year, the proportion of comments that are informative increases to 81.0%. **Therefore, later comments are more likely to be informative.** As an example, in Jan 2016 an online tutorial was posted in an accepted answer about installing and switching PHP versions. More than one year later (i.e., in May 2017), a comment[18] pointed out that the tutorial was moved. In another example, a comment[19] mentioned more than one year later (i.e., April 2017) that "*this no longer works ... changing that class selector fixes it*" under an accepted answer that was originally posted in January 2016. Hence, we encourage users to carefully read late arriving comments because of their likelihood to point out incorrect or updated information (e.g., an answer is obsolete [1]).

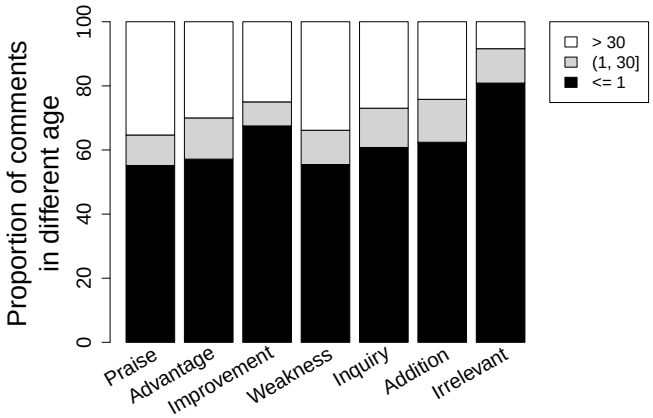

Fig. 7: The proportion of comments within different ranges of commenting time (in days) for each comment type.

The majority of commenting activities occur after their associated answers is accepted. Among all the accepted answers with comments, 77.1% of the comments are posted after their associated answer is accepted. Fig. 8 shows the proportion of the tagged comments posted before and after their associated accepted answer. Even though askers have selected the accepted answers and the community tends to consider accepted answers as the "best" answers, **the community does not necessarily stop discussing these answers through commenting**. We suggest answer seekers to carefully read through the commenting-thread (i.e., the flattened list of all the comments that are associated with an answer after clicking "show N more comments" instead of only the top 5 comments that are displayed by default), even in the accepted answers. **Even though an answer is accepted, it does not necessarily mean that it is proven to be the "best", and any comment that is associated with this answer can potentially enhance the answer itself.** For

18. https://www.stackoverflow.com/posts/comments/75190051
19. https://www.stackoverflow.com/posts/comments/73451366

example, in Fig. 1, the answer was accepted in July 2010, and the comment that made a correction to the answer was actually posted in March 2013, i.e., after more than two years since the answer was accepted. So far, this comment has gained a score that is 3 times higher than the score of its associated answer. Zhang et al. [1] also observed that "30.7% of the (studied) obsolete accepted answers got updated after being noted as obsolete (by a comment that is associated with the obsolete answer)".

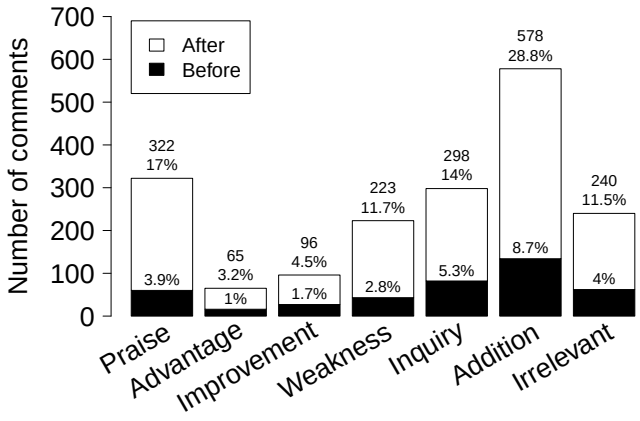

Fig. 8: The proportion of comments that are posted before and after their associated answer is accepted for each comment type. Note that the total number of the comments shown in this figure is less than 2,000 since some manually studied comments were posted under non-accepted answers.

Answers are rarely updated after comments are posted, indicating that comments are rarely integrated back into answers – thus, users have to carefully read the comments. In the 11.4 million answers with comments, 61.9% (i.e., 7.1 million) of the answers have never been edited since their creation. Only 14.1% (i.e., 1.6 million) of the answers are edited after any comment. Note that 14.1% is an upper bound, since the edits in answers may not be related to the posted comments. Although the majority of comments are informative based on our findings in Section 4.1, comments are rarely integrated back into their associated answers. Therefore, we suggest the Stack Overflow team to encourage users to maintain answers with badges and reputation points, e.g., rewarding users who actively update answers by leveraging their associated comments.

The acceptance of an answer is not the end of commenting activities; instead, the majority of commenting activities occur after the acceptance of an answer. Generally, more than half of the comments are fast responses occurring within one day of the answer creation. Comments of type *advantage* and *weakness* are more likely to be posted later than other informative comments, and later comments tend to be more informative. Even though most comments provide useful information, they are rarely integrated back into answers.

4.3 RQ3: What types of users participate in the commenting activities?

Motivation: Stack Overflow sets restrictions on who can post comments. Namely, any user with at least 50 reputation points, the owner of the answer, and the owner of the question thread can post comments. To better organize commenting-threads, it is important to understand what types of users participate in these activities. Part of the reasons why organizing such a large collection of comments is challenging, is that comments are posted in a crowdsourced manner by different users. Therefore, in this RQ, we wish to understand how commenters with different roles (e.g., asker and answerers) are involved in the commenting-threads.

Approach: Based on the role of a commenter in the entire question thread, we categorize commenters into one of the following three groups:

1) **Asker**: the user who posted the question;
2) **Answerer**: the user who posted the answer;
3) **Outsider**: the user who belongs to neither of the two above-mentioned roles in that question thread.

We refer to an asker or answerer who is involved in the question thread (groups 1 & 2) as an *insider* (since they were involved earlier in the question answering process). The role of a commenter is assigned using the priority: asker $>$ answerer $>$ outsider. For example, if a user has multiple roles, such as an asker and an answerer, we consider the user as an asker.

Furthermore, we analyze how the roles of commenters are correlated with the comment types and the temporal dynamics of comments, in terms of the commenting time and whether commenting occurs before or after an answer acceptance.

To find out how experienced are the commenters when they post comments, we analyze a user's reputation within each comment type. Since a user's reputation changes over time, we crawl the daily activities of a user from their user profile webpages and calculate the reputation points of the user when he/she posted a comment. We analyze the relationship between the types of comments and the reputation points of users when they posted the comments.

Results: In general, users are actively posting comments. All the 32.3 million comments are posted by 1.9 million users, compared with 1.7 million users who post 22.7 million answers. On average, among all the answers and their associated comments in each question thread, the ratio of the number of commenters to answerers is 1.675:1, that is, there are 67.5% more commenters than answerers per question thread. The median of the ratio of the number of commenters to answerers is 1.75:1. As of April 16, 2019, the top commenter (i.e., Jon Skeet) has posted 43,876 comments that are associated with 24,568 answers. Stack Overflow uses badges to encourage users to leave comments under posts. As a result, 881,649 *Commentator* badges (i.e., to award users who leave 10 comments) and 10,232 *Pundit* badges (i.e., to award users who leave 10 comments with a score of 5 or more) are rewarded to users[20].

The majority of comments are posted by insiders within one day since the creation of answers, while outsiders are more likely to post comments after one

20. Data obtained on April 11, 2019 from https://stackoverflow.com/help/badges.

month. Among all the comments that are associated with answers, 45.2% are posted by askers, and 31.1% are posted by answerers. 62.3% of the commenters are users who post under their own answers or questions (i.e., users with lower than 50 reputation points). As shown in Fig. 9, within one month, the majority of comments are posted by either askers or answerers (i.e., insiders). The dynamics of commenting activities are similar to the dynamics of answering activities on Stack Overflow, i.e., most questions get their accepted answers in half an hour [6]. However, after one month, the majority of comments are posted by outsiders (i.e., the user did not post the question nor the answer, before posting the comment). In Section 4.2, we find that later comments are more informative; therefore, after one month, outsiders start to play an important role in maintaining the associated answers through commenting activities.

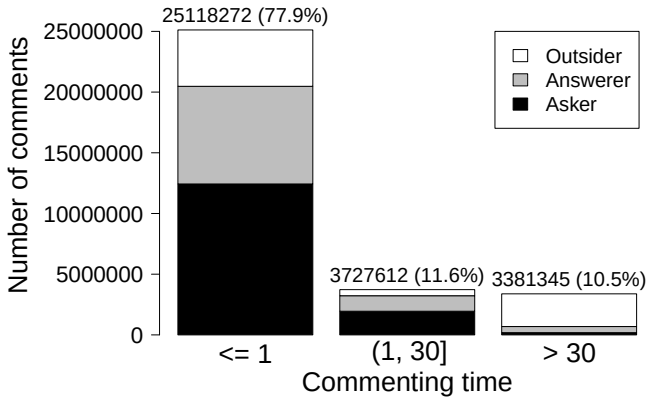

Fig. 9: The number and proportion of comments that were posted by different user roles in different ranges of commenting time.

These results suggest that the maintenance of crowdsourced knowledge is a long-term task, and later activities should not be neglected. While outsiders do not contribute to the discussion in earlier stages, they are significantly involved later on. In addition, we observe that askers, answers, and outsiders are all involved in the commenting activities, both before and after the acceptance of an answer, as shown in Fig. 10.

Askers mainly post comments that belong to type *praise*, *inquiry*, and *weakness*. Answerers mainly post comments in type *advantage*, *improvement*, *addition*, and *irrelevant*. The proportion of commenter roles in each comment type is shown in Fig. 11. We notice that a significant proportion of comments of type *advantage*, *improvement*, and *weakness* are posted by outsiders, although these outsiders never participate in the entire question thread before posting the comment. Based on the above-mentioned observations, Stack Overflow can design a better channel for askers to appreciate answers. **A new praise channel instead of commenting can separate praising activities from commenting activities that can enhance the value of an answer.** The praise channel helps make a commenting-thread less crowded with irrelevant comments. As examples, both GitHub reactions[21] and Basecamp boosts[22] are such designs to channel praise comments. Stack Overflow can also provide the associated askers and answerers with alternative channels instead of posting comments of type *praise* and *inquiry*, such as praising by sending iconic expressions to answerers or other commenters instead of praising in comments. Only the praised users are concerned with such praising activities, while the commenting area can be a place for the community to discuss the answer with praising content hidden or removed.

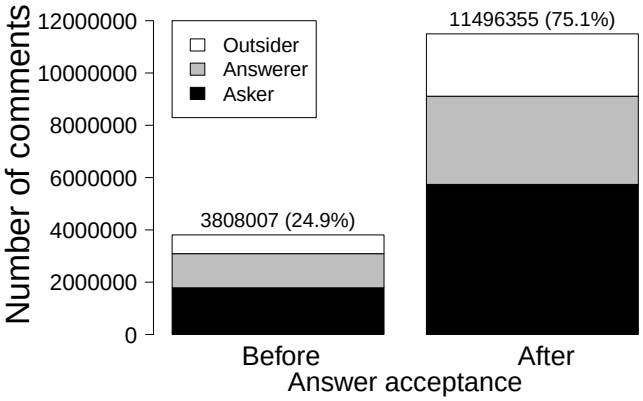

Fig. 10: The number and proportion of comments that were posted by different user roles before and after the answer acceptance.

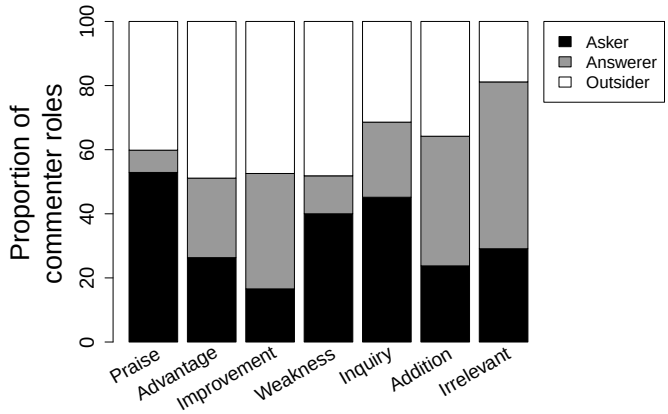

Fig. 11: Proportion of commenter roles in each comment type.

Among users who post informative comments, inexperienced users (i.e., ones with lower reputation) tend to raise limitations and concerns by posting comments of type *weakness* and *inquiry*, while experienced users (i.e., ones with higher reputation) tend to enhance the answer with their comments by posting comments of type *advantage*, *improvement*, and *addition*. Fig. 12 shows the distribution of reputation points for users who post in each comment type. Even though comments of type *weakness* and *inquiry* are often posted by users with lower reputation points (with a median value of 465 and 423) than comments

21. https://developer.github.com/v3/reactions/
22. https://3.basecamp-help.com/article/391-boosts

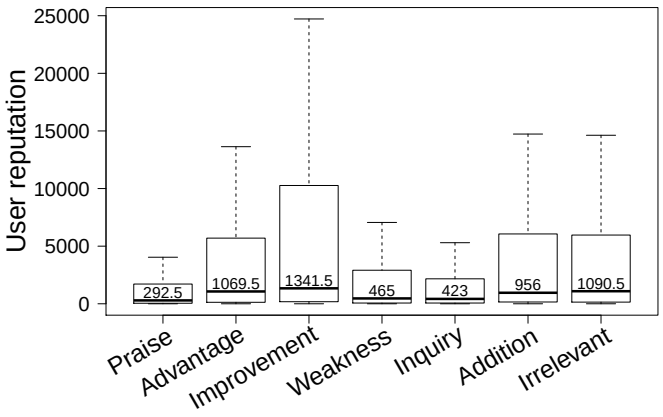

Fig. 12: The distribution of user reputation points in each comment type (the median value is shown inside each box).

of type *advantage*, *improvement*, and *addition* (with a median value of 1,069.5, 1,341.5, and 956, respectively), all these users are actively contributing to enhance the associated answers. To further test if these differences are statistically significant, we ran the Mann-Whitney U test between the distribution of user reputation points for the comments of type *weakness* and each one of the three other types (i.e., *advantage*, *improvement*, and *addition*). We find that the difference is statistically significant with p-value $< 0.05/3$ (adjusted with a Bonferroni correction) in all three cases. The reputation points of users who post the comments of type *inquiry* is also statistically significantly lower than each one of the three other comment types that enhance their associated answers (i.e., *advantage*, *improvement*, and *addition*) with the p-values all below $0.05/3$ (adjusted with a Bonferroni correction).

Among users who post uninformative comments, inexperienced users (i.e., with median reputation points of 292.5) tend to post comments of type *praise*. These commenters are probably not familiar with Stack Overflow, and simply express their appreciation through commenting instead of upvoting or accepting answers. **Experienced users (i.e., with median reputation points of 1,090.5) tend to post comments of type *irrelevant*.** Even though these users have reputation as high as users who post informative comments, they do not necessarily enhance the associated answers by commenting.

Furthermore, we group users by their reputation points. Fig. 13 shows the proportion of comments in different types that are posted by different user groups. The reputation thresholds among different user groups are defined by Stack Overflow[23]. We find that users with higher reputation points are more likely to post a lower proportion of comments of type *praise*, *weakness*, and *inquiry*, and are more likely to post a higher proportion of comments of type *advantage*, *improvement*, *addition*, and *irrelevant*. Such users with higher reputation points are probably more aware of the community rules; thus, posting fewer comments to praise an answer or make an additional inquiry. Users with higher reputation points also contribute to the crowdsourced knowledge sharing through the frequent posting of comments of type *advantage*, *improvement*, and *addition*. Surprisingly, users with higher reputation points post fewer comments to point out weaknesses and more irrelevant comments.

23. https://stackoverflow.com/help/privileges?tab=milestone

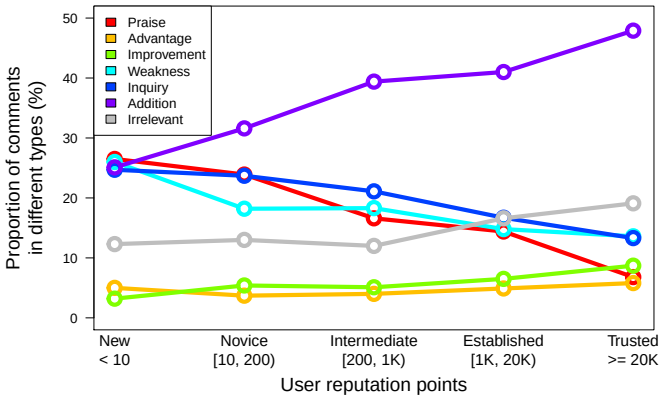

Fig. 13: The proportion of comments in different types among different user groups.

Currently, commenting activities do not reward any reputation point on Stack Overflow. Even if a comment is extremely helpful and gets a large number of upvotes, the commenter will not gain any reputation points. In the example shown in Fig. 1, although the comment got 1,135 scores compared to a score of 376 for the answer, the commenter gained **no** reputation points while the answerer gained 3,760 reputation points (i.e., 10 reputation points for each one of the 376 upvotes). This commenter contributed to the maintenance of the crowdsourced knowledge and is recognized by the community (i.e., through the comment score), but he did not receive any reward. Although there exists 2 badges (i.e., *Commentator*, which is given to commenters who leave 10 comments, and *Pundit*, which is given to commenters who leave 10 comments with score of 5 or more) are related to commenting activities on Stack Overflow, only the users who reach these specific criteria can receive these badges, regardless of the usefulness and importance of any of their comments. These two badges are also designed for comments posted under both questions and answers; therefore, they are not directly designed for encouraging users to maintain Stack Overflow answers (note that 2,000 reputation points are required to maintain an answer by directly editing). On the other hand, the upvoting of an answer by other users directly adds reputation points to the answerer.

Users are highly involved in commenting. Askers mainly post comments that belong to type *praise*, *inquiry*, and *weakness*, while answerers mainly post comments of type *advantage*, *improvement*, and *addition*. Insiders post the majority of comments within one month, while outsiders are more likely to post comments after one month. Among informative comments, inexperienced users tend to raise limitations and concerns while experienced users tend to enhance the answer by commenting. Among uninformative comments, inexperienced users tend to praise the answer while experienced users tend to post irrelevant information.

5 Implications of our Findings

5.1 Implications for Stack Overflow and Users

Although existing answers can be revised and new answers can be created in their associated question threads for updating existing knowledge, it is unclear how effectively do users maintain answers. In addition, the evolution of the underlying programming languages, APIs, and other software artifacts makes it challenging to keep the 22.7 million Stack Overflow answers up to date, i.e., it is challenging to evaluate the answer quality in the long term. On the other hand, comments provide additional observations to their associated answers, such as answer obsolescence [1] and security flaws in answers[24]. Under these scenarios, users who post these informative comments play an important role in maintaining the existing crowdsourced knowledge by observing and even addressing issues in answers. Therefore, these commenting activities can improve the long term value of their associated answers.

Based on our findings, we encourage Stack Overflow designers & engineers to focus on how to more effectively maintain the crowdsourced knowledge on Stack Overflow by leveraging the large collection of comments. We note that the proportion of answers that currently have been updated based on the rich content in comments is low. We provide below some implications for Stack Overflow and users based on our findings:

1) Since informative comments can significantly enhance their associated answers, we propose that these commenters are rewarded with reputation points, thus motivating the maintenance of crowdsourced knowledge. 4.4 million (i.e., 38.9%) of the answers with comments have a comment with an equal or higher score than the answer itself. However, under the current reputation system, a commenter does not gain any reputation points while an answerer gains 10 reputation points from each upvote.
2) Stack Overflow should encourage users to maintain answers, e.g., by rewarding users who leverage comments to update answers with badges and reputation points. As Stack Overflow and the knowledge within it age (Stack Overflow is over 10 years old today), many answers on Stack Overflow are likely to become outdated relative to the latest technologies. We already observed many answers that are not updated to reflect informative comments on these answers. Therefore, knowledge maintenance should be actively encouraged. For example, a checkbox of "answer maintenance" can be provided to users who post comments to indicate that the posted comments can be potentially used to maintain the answer, and a review queue can be added for these types of comments. When a user posts a comment which could be used to maintain the answer, the user can check the "answer maintenance" checkbox then this comment will be added into a queue for the community to review. If the community agrees with the comment, the comment could be labeled as "answer maintenance" to indicate its value. If these comments that serve the purpose of maintaining answers get approval after the review process, they can be highlighted and their corresponding users can be awarded through the gamification mechanism (e.g., through badges as done for answer editing badges [4]).
3) Users can tag their comments based on our existing comment types. With tagged comments, a better organization scheme can be implemented to display comments, thus leveraging the massive collection of informative comments for the purpose of both answer maintenance and information retrieval. In addition, an automated classifier can be developed to identify informative comments and comments of different types. The observed characteristics from our study of the temporal dynamics of commenting activities can provide insights for future work to build such an automated classifier.
4) Comments of type *praise* exist while they do not improve the quality of an answer. A classifier can be implemented to detect comments of type *praise*. Users can be suggested to upvote an answer instead of posting a comment. By removing these comments of type *praise*, users can retrieve informative comments more effectively, which eventually assists them in solving their issues.
5) Unrecommended uses of comments can be flagged to help users follow Stack Overflow's guidelines. Comments that suggest corrections, answer a question, or relay a compliment can be automatically detected, and proper actions can be suggested to these commenters. A classifier to automatically identify such unrecommended uses of comments can be built, or individual classifiers can be built to tackle each unrecommended case. As a result, unnecessary comments can be deleted and users can retrieve informative comments more effectively. Similarly, a classifier to identify informative comments can be built to effectively assist users in retrieving relevant information from comments. Our findings can be leveraged by future work for comment classification. For example, we find in Fig. 13 that users with higher reputation points post a higher proportion of certain types of informative comments (e.g., *addition* and *improvement*). The reputation of a commenter may be used as a potential metric to identify informative comments. Furthermore, Stack Overflow can provide notifications to users about posting potentially uninformative comments, thus the overall informativeness of comments throughout Stack Overflow can be further enhanced.
6) Without any active organizing effort, the best suggestion so far for users is to read every single comment carefully, regardless of whether it is displayed or not. In particular, a reader is suggested to read later comments since they are more likely to be informative. Finally, to gain a closer look at users' opinions on Stack Overflow comments, we conduct a preliminary user survey to ask 22 participants the following question: "*Do you read comments when you use Stack Overflow?*". Out of the 22 responses, 9 participants read comments occasionally and 1 participant never read comments. Our study shows that 45.4% of the participants do not actively read comments. Hence, based on our findings, we encourage users to read comments carefully since

24. https://www.attackflow.com/Blog/StackOverflow

the majority (i.e., 75.6%) of comments provide a diverse variety of information that enhances their associated answers.

Note that any gamification mechanism on Stack Overflow may have adverse side effects as noted by Wang et al. in recent work on the use of badges in Stack Overflow [4]. For example, awarding reputation points for commenting activities could lead to an increase in the number of uninformative comments by users who attempt to fish for reputation points by posting comments. Future studies are encouraged to study the impact of gamification mechanism on the user participation and its side effect to have a better balance.

5.2 Implications for Researchers

Another implication of our study is for researchers. Since 2009, many research efforts continue to leverage the Stack Overflow dataset. The majority of the studies only leveraged the information related to questions and answers. There exists a limited number of prior studies that leveraged the information from comments. For example, Zou et al. analyzed both posts and comments to investigate non-functional requirements on Stack Overflow [7]. Castelle evaluated the classification models of abusive language from Stack Overflow comments [8]. We encourage future research to leverage the 32.3 million comments that are associated with answers to actively support maintenance efforts of such crowdsourced knowledge. We observe that answers can be updated through the leaving of informative comments on these answers. Therefore, reviewing comments is recommended when analyzing Stack Overflow answers. Furthermore, researchers can leverage such rich and informative comments to enhance various software engineering tasks, e.g., API documentation enhancement [9] and question answering bot [10].

In addition, our paper is the first work to empirically study the types of information in comments. In comparison, Poché et al. found that 30% of the comments on YouTube coding tutorials are informative [11], while Chen et al. found that 35% of app reviews from Google Play are informative [12]. We observe that the majority (∼76%) of these comments are informative and enhance answers from a diverse range of perspectives. Future studies may propose data-driven solutions for retrieving informative comments to either identify or summarize such comments in an automated manner. Future research can leverage approaches from the machine learning and natural language processing communities to automatically identify the comment types/subtypes that we identified. The identified comments may assist developers with the reading of Stack Overflow posts or assist researchers to better leverage the information in comments.

6 Threats to Validity

External validity. Threats to external validity relate to the generalizability of our findings. The number of comments is large and it is impossible to study all of the comments in our qualitative study. In order to minimize the bias, we randomly sampled 3,000 statistically representative comments, giving us a confidence level of 99% and a confidence interval of 2.4%. In this study, we focus on Stack Overflow, which is one of the most popular Q&A websites for developers, hence, our results may not generalize to other Q&A websites. To alleviate this threat, more Q&A websites should be studied in the future. Furthermore, in this study, we analyzed comments that are associated with answers. The comments that are associated with questions can also be informative, and thus, contribute to the crowdsourced knowledge sharing on Stack Overflow. Future research should investigate questions comments and explore how such comments enhance the question answering activities on Stack Overflow.

Internal validity. Threats to internal validity are related to experimenter errors and bias. Our study involved qualitative studies which were performed by humans. Bias may be introduced. To reduce the bias of our analysis, each comment is labeled by two of the authors individually and discrepancies are discussed until a consensus is reached. We measured the level of the inter-rate agreement in our qualitative study, and the agreement value is substantial (i.e., 0.86 and 0.90 at the subtype and type level, respectively) even before the consensus is reached.

7 Related Work

7.1 Knowledge sharing and management for Stack Overflow

Stack Overflow is a popular online community for developers to provide solutions and exchange ideas. Programming knowledge is embedded in millions of questions and their answers aiming to solve individual programming issues. Thus, a considerable number of studies have been done on the Stack Overflow dataset to gain a deeper understanding of the crowdsourced knowledge sharing and management among developers [1], [4]–[6], [13]–[31]. For example, Treude et al. surveyed the self-explanatory nature of code fragments on Stack Overflow, and identified the main causes of code understanding challenges [29]. Calefato et al. investigated factors to increase the chances of answer acceptance on Stack Overflow [13]. They suggested question answering is a two-phase activity, where answerers should not only write an answer, but also avoid any negative attitude towards askers in comments. Dalip et al. proposed an approach to rank answers based on the feedback given to answers [14]. They observed that both user and review features are important to assess the quality of answers. Choetkiertikul et al. proposed approaches to route questions to specific answerers using both feature-based and social network approaches [15]. Their approach can enhance the knowledge exchanging in the Stack Overflow community. Xu et al. proposed answer summarization by leveraging both relevant questions and the usefulness & diversity of answers [17]. Other researches have worked on enhancing the content management of answers on Stack Overflow. For example, Srba et al. analyzed the evolution of activities on Stack Overflow and found that low-quality content and certain types of users (e.g., newcomers and reputation collectors) are closely correlated with the long-term sustainability of Stack Overflow [19]. Fischer et al. found that 30.9% of

Android code snippets were insecure in Android-related answers [18]. Zhang et al. found that when an answer was observed as obsolete, only 20.5% of such answers are ever updated [1]. An et al. analyzed 399 Android apps and observed 1,279 cases of potential license violations among reused code both from and to Stack Overflow [30]. Ragkhitwetsagul et al. analyzed Java code snippets on Stack Overflow and found that 153 clones were copied to Stack Overflow with 66% (i.e., 100) being outdated [31].

In this study, we leverage the large collection of comments that are associated with answers in order to explore how they actually add value to existing answers. We find that although comments are informative, they are rarely integrated back into their associated answers. We encourage future research to enhance the quality of answers by utilizing the knowledge embedded in these extended discussions among commenters.

7.2 User feedback through online commenting

Software systems are not isolated from their users. User feedback, although usually in a form of unstructured text, directly reflects how users are satisfied with the current system. A large amount of online comments are studied by prior research [1], [11], [32]–[39]. For example, Galvis et al. [32] analyzed user comments in the Google Play Store. They applied topic modeling to extract topics related to requirements changes in the application development process. Park et al. [33] developed a system to help comment moderators identify high quality comments on online news sites. Their approach enabled the identification of high quality content at a large scale in online journalistic systems. Poché et al. [11] analyzed user comments from YouTube coding tutorial videos, and classified useful feedback that requires further action from the content creator. Additionally, comments on Stack Overflow have been leveraged in prior work [1], [7], [8], [34]. For example, Cleary et al. [34] manually labeled a large number of Stack Overflow comments for their friendliness in order to gain more insights about negative comments in developer communities. They found that the most unfriendly comments use a constrained vocabulary, and this high degree of pattern repetition can be leveraged for automatically identifying unfriendly comments. Zhang et al. leveraged comments that observed the obsolescence of answers [1]. They applied a heuristic-based keyword search approach to identify obsolete answers from their associated comments with an accuracy of 75%.

Instead of exploring certain aspects of comments on Stack Overflow, we focus on the whole collection of 32.3 million comments that are associated with answers, and conduct an empirical study to analyze this dataset and investigate the characteristics of comments. We wish to extract in-depth insights so that informative comments can be more effectively integrated into the crowdsourced question answering process.

8 Conclusion

In this paper, we investigate 32.3 million comments that are associated with answers on Stack Overflow. Since 2009, users create more comments than answers on a yearly basis. 23% (i.e., 2.6 million) of the answers with comments even have a commenting-thread longer than the actual answer, indicating the richness of information in comments.

Our empirical study provides an in-depth understanding of the commenting activities on Stack Overflow. We identify various types of comments and find that the majority of comments are informative as they enhance answers from a diverse range of perspectives. We also characterize the commenting activities in terms of time and user roles. We find that comments are rarely integrated back into their associated answers. Insiders (i.e., askers and answerers) post the majority of comments within one day, while outsiders (i.e., users with no earlier activity within a question thread) post the majority of comments after one month. These outsiders also post informative comments.

Our analysis can be leveraged to create alternative channels for askers and answerers to request detailed information and receive compliments, respectively. The informative comments can also be further utilized to actively maintain their associated answers and improve their presentation. Our findings can be leveraged for crowdsourced knowledge maintenance and organization.

References

[1] H. Zhang, S. Wang, T. P. Chen, Y. Zou, and A. E. Hassan, "An empirical study of obsolete answers on Stack Overflow," *IEEE Transactions on Software Engineering*, 2019.

[2] Stack Overflow, "Privileges - comment everywhere," https://stackoverflow.com/help/privileges/comment, 2019, online; accessed May 16 2019.

[3] Z. Reich, "User comments," in *Participatory Journalism*. John Wiley & Sons, Ltd, 2011, ch. 6, pp. 96–117.

[4] S. Wang, T.-H. P. Chen, and A. E. Hassan, "How do users revise answers on technical Q&A websites? a case study on stack overflow," *IEEE Transactions on Software Engineering*, 2018.

[5] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: A case study of Stack Overflow," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12, 2012, pp. 850–858.

[6] S. Wang, T.-H. Chen, and A. E. Hassan, "Understanding the factors for fast answers in technical Q&A websites," *Empirical Software Engineering*, vol. 23, no. 3, pp. 1552–1593, Jun 2018.

[7] J. Zou, L. Xu, W. Guo, M. Yan, D. Yang, and X. Zhang, "Which non-functional requirements do developers focus on? An empirical study on Stack Overflow using topic analysis," in *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, May 2015, pp. 446–449.

[8] M. Castelle, "The linguistic ideologies of deep abusive language classification," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 160–170.

[9] C. Treude and M. P. Robillard, "Augmenting API documentation with insights from Stack Overflow," in *Proceedings of the 38th International Conference on Software Engineering*, ser. ICSE '16. New York, NY, USA: ACM, 2016, pp. 392–403. [Online]. Available: http://doi.acm.org/10.1145/2884781.2884800

[10] Y. Tian, F. Thung, A. Sharma, and D. Lo, "APIBot: Question answering bot for API documentation," in *Proceedings of the 32Nd IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE 2017. Piscataway, NJ, USA: IEEE Press, 2017, pp. 153–158. [Online]. Available: http://dl.acm.org/citation.cfm?id=3155562.3155585

[11] E. Poché, N. Jha, G. Williams, J. Staten, M. Vesper, and A. Mahmoud, "Analyzing user comments on YouTube coding tutorial videos," in *Proceedings of the 25th International Conference on Program Comprehension*, ser. ICPC '17, 2017, pp. 196–206.

[12] N. Chen, J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang, "AR-miner: Mining informative reviews for developers from mobile app marketplace," in *Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE 2014. New York, NY, USA: ACM, 2014, pp. 767–778.

[13] F. Calefato, F. Lanubile, M. C. Marasciulo, and N. Novielli, "Mining successful answers in Stack Overflow," in *Proceedings of the 12th Working Conference on Mining Software Repositories*, ser. MSR '15, 2015, pp. 430–433.

[14] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Exploiting user feedback to learn to rank answers in Q&A forums: A case study with Stack Overflow," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13, 2013, pp. 543–552.

[15] M. Choetkiertikul, D. Avery, H. K. Dam, T. Tran, and A. Ghose, "Who will answer my question on Stack Overflow?" in *2015 24th Australasian Software Engineering Conference*. IEEE, 2015, pp. 155–164.

[16] N. Meng, S. Nagy, D. D. Yao, W. Zhuang, and G. A. Argoty, "Secure coding practices in Java: Challenges and vulnerabilities," in *Proceedings of the 40th International Conference on Software Engineering*, ser. ICSE '18, 2018, pp. 372–383.

[17] B. Xu, Z. Xing, X. Xia, and D. Lo, "Answerbot: Automated generation of answer summary to developers' technical questions," in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2017, pp. 706–716.

[18] F. Fischer, K. Bttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl, "Stack Overflow considered harmful? The impact of copy & paste on Android application security," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 121–136.

[19] I. Srba and M. Bielikova, "Why is Stack Overflow failing? preserving sustainability in community question answering," *IEEE Software*, vol. 33, no. 4, pp. 80–89, July 2016.

[20] S. Wang, D. Lo, and L. Jiang, "An empirical study on developer interactions in StackOverflow," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ser. SAC '13, 2013, pp. 1019–1024.

[21] B. Xu, D. Ye, Z. Xing, X. Xia, G. Chen, and S. Li, "Predicting semantically linkable knowledge in developer online forums via convolutional neural network," in *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE 2016, 2016, pp. 51–62.

[22] S. Wang, D. Lo, B. Vasilescu, and A. Serebrenik, "Entagrec ++: An enhanced tag recommendation system for software information sites," *Empirical Software Engineering*, vol. 23, no. 2, pp. 800–832, 2018.

[23] Y. Wu, S. Wang, C.-P. Bezemer, and K. Inoue, "How do developers utilize source code from stack overflow?" *Empirical Software Engineering*, pp. 1–37, 2018.

[24] A. Bacchelli, L. Ponzanelli, and M. Lanza, "Harnessing Stack Overflow for the IDE," in *Proceedings of the Third International Workshop on Recommendation Systems for Software Engineering*, ser. RSSE '12, 2012, pp. 26–30.

[25] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider, "Answering questions about unanswered questions of Stack Overflow," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR '13, 2013, pp. 97–100.

[26] Q. Tian, P. Zhang, and B. Li, "Towards predicting the best answers in community-based question-answering services," in *Proceedings of the 7th International Conference on Weblogs and Social Media*, ser. ICWSM 2013, 2013, pp. 725–728.

[27] S. Chang and A. Pal, "Routing questions for collaborative answering in community question answering," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13, 2013, pp. 494–501.

[28] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov, "How social Q&A sites are changing knowledge sharing in open source software communities," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW '14, 2014, pp. 342–354.

[29] C. Treude and M. P. Robillard, "Understanding Stack Overflow code fragments," in *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 2017, pp. 509–513.

[30] L. An, O. Mlouki, F. Khomh, and G. Antoniol, "Stack Overflow: a code laundering platform?" in *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2017, pp. 283–293.

[31] C. Ragkhitwetsagul, J. Krinke, M. Paixao, G. Bianco, and R. Oliveto, "Toxic code snippets on stack overflow," *IEEE Transactions on Software Engineering*, 2019.

[32] L. V. Galvis Carreño and K. Winbladh, "Analysis of user comments: an approach for software requirements evolution," in *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013, pp. 582–591.

[33] D. Park, S. Sachar, N. Diakopoulos, and N. Elmqvist, "Supporting comment moderators in identifying high quality online news comments," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 1114–1125.

[34] B. Cleary, C. Gómez, M.-A. Storey, L. Singer, and C. Treude, "Analyzing the friendliness of exchanges in an online software developer community," in *2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*. IEEE, 2013, pp. 159–160.

[35] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, and N. Sadeh, "Why people hate your app: Making sense of user feedback in a mobile app store," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1276–1284.

[36] E. Shmueli, A. Kagian, Y. Koren, and R. Lempel, "Care to comment? recommendations for commenting on news stories," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 429–438.

[37] L. Cen, L. Si, N. Li, and H. Jin, "User comment analysis for android apps and CSPI detection with comment expansion," in *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security (PIR 2014)*, p. 25.

[38] R. Laiola Guimarães, P. Cesar, and D. C. Bulterman, "Let me comment on your video: Supporting personalized end-user comments within third-party online videos," in *Proceedings of the 18th Brazilian symposium on Multimedia and the web*. ACM, 2012, pp. 253–260.

[39] B.-C. Chen, J. Guo, B. Tseng, and J. Yang, "User reputation in a comment rating environment," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 159–167.

Haoxiang Zhang Haoxiang Zhang is currently working toward a PhD degree in the School of Computing at Queen's University, Canada. He received his PhD degree in Physics from Lehigh University, Bethlehem, Pennsylvania in 2013. His research interests include machine learning in software analytics, empirical software engineering, and mining software repositories. More information at: https://haoxianghz.github.io/.

Shaowei Wang Shaowei Wang is an assistant professor in the Department of Computer Science and Engineering at Mississippi State University. Before joining MSU, he was a postdoctoral fellow in the Software Analysis and Intelligence Lab (SAIL) at Queen's University, Canada. He obtained his Ph.D. from Singapore Management University and his BSc from Zhejiang University. His research interests include software engineering, machine learning, data analytics for software engineering, automated debugging, and secure software development. He is one of four recipients of the 2018 distinguished reviewer award for the Springer EMSE (SE's highest impact journal). More information at: https://sites.google.com/site/wswshaoweiwang/.

Tse-Hsun (Peter) Chen Tse-Hsun (Peter) Chen is an Assistant Professor in the Department of Computer Science and Software Engineering at Concordia University, Montreal, Canada. He leads the Software PErformance, Analysis, and Reliability (SPEAR) Lab, which focuses on conducting research on performance engineering, program analysis, log analysis, production debugging, and mining software repositories. His work has been published in flagship conferences and journals such as ICSE, FSE, TSE, EMSE, and MSR. He serves regularly as a program committee member of international conferences in the field of software engineering, such as ASE, ICSME, SANER, and ICPC, and he is a regular reviewer for software engineering journals such as JSS, EMSE, and TSE. Dr. Chen obtained his BSc from the University of British Columbia, and MSc and PhD from Queen's University. Besides his academic career, Dr. Chen also worked as a software performance engineer at BlackBerry for over four years. Early tools developed by Dr. Chen were integrated into industrial practice for ensuring the quality of large-scale enterprise systems. More information at: http://petertsehsun.github.io/.

Ahmed E. Hassan Ahmed E. Hassan is an IEEE Fellow, an ACM SIGSOFT Influential Educator, an NSERC Steacie Fellow, the Canada Research Chair (CRC) in Software Analytics, and the NSERC/BlackBerry Software Engineering Chair at the School of Computing at Queen's University, Canada. His research interests include mining software repositories, empirical software engineering, load testing, and log mining. He received a PhD in Computer Science from the University of Waterloo. He spearheaded the creation of the Mining Software Repositories (MSR) conference and its research community. He also serves/d on the editorial boards of IEEE Transactions on Software Engineering, Springer Journal of Empirical Software Engineering, and PeerJ Computer Science. Contact ahmed@cs.queensu.ca. More information at: http://sail.cs.queensu.ca/.