Attribution Required: Stack Overflow Code Snippets in GitHub Projects

Sebastian Baltes, Richard Kiefer, and Stephan Diehl
Software Engineering Group
University of Trier
Trier, Germany
research@sbaltes.com, diehl@uni-trier.de

***Abstract*—Stack Overflow (SO) is the largest Q&A website for developers, providing a huge amount of copyable code snippets. Using these snippets raises various maintenance and legal issues. The SO license requires attribution, i.e., referencing the original question or answer, and requires derived work to adopt a compatible license. While there is a heated debate on SO's license model for code snippets and the required attribution, little is known about the extent to which snippets are copied from SO without proper attribution. In this paper, we present the research design and summarized results of an empirical study analyzing attributed and unattributed usages of SO code snippets in GitHub projects. On average, 3.22% of all analyzed repositories and 7.33% of the popular ones contained a reference to SO. Further, we found that developers rather refer to the whole thread on SO than to a specific answer. For Java, at least two thirds of the copied snippets were not attributed.**

***Keywords*-empirical study; code snippets; copy-and-paste programming; licensing; stack overflow; github; survey**

I. Introduction

Stack Overflow is the largest Question and Answer (Q&A) website for software developers. As of February 2017, its public data dump [1] lists 11.5 million answered questions and 6.7 million registered users. Many of the answers contain code snippets together with explanations [2]. The availability of this huge amount of code snippets lead to changes in software developers' behavior: Nowadays, they regularly face the "build or borrow" question [3]: Should they try to understand and solve an issue on their own or just copy and adapt a solution from Stack Overflow (SO)? Assuming that developers also copy and paste snippets from SO without trying to thoroughly understand them, *maintenance issues* may arise. For instance, it will later be difficult for developers to refactor or debug code that they did not write themselves. Moreover, if no link to the SO question or answer is added to the copied code, it is not possible to check the SO thread for a corrected or improved solution in case problems occur. Beside these maintainability implications, copying and pasting code from SO may also lead to *licensing issues*: All SO content is licensed under the Creative Commons Attribution-ShareAlike 3.0 license (CC BY-SA 3.0). This license allows to share and adapt the content, but requires attribution and demands contributions based on the content to be published under a compatible license. Licensing issues of source code posted on SO have been controversially discussed on different sites of the Stack Exchange network [4], [5], [6], but attempts to change the license failed [7], [8]. Besides attribution, CC BY-SA requires derived work to use a compatible license. However, CC licenses are not common for software [9] and there is currently no non-CC license that Creative Commons considers compatible [10]. This makes the usage of code snippets from SO problematic in terms of licensing conflicts. With more than 33 million repositories, GitHub (GH) is one of the most popular code hosting platforms. To the best of our knowledge, there is currently no empirical evidence on how common it is to copy&paste code from SO into GH projects, and in particular, how often content from SO is attributed as required by the license. In the following, we present the research design and summarized results of a first thorough analysis on the usage and attribution of SO code snippets in public software projects hosted on GH. To complement our results, we conducted two surveys with software developers on their attribution practice and their awareness regarding the licensing of code from SO posts. With our research, we want to take a first step towards quantifying the impact of SO on open-source software development. We also want to make developers aware of their attribution practice and possible licensing and maintainability issues.

II. Research Design

The overall goal of our research was to analyze the *usage* and *attribution* of code snippets from Stack Overflow in GitHub projects. By usage we mean copying (and possibly adapting) the code snippet from an answer on SO and then pasting it into a public GH project. Figure 1 depicts the high-level research design. We started our research with a preliminary study to get first insights into if and how developers copy code snippets from SO. To validate the preliminary result that many developers do not attribute code copied from SO, we utilized six data sources: the *GitHub* [11] and *Stack Exchange* APIs [12], the *GHTorrent* [13] and *Stack Overflow* [14] data dumps, and the *BigQuery GitHub* [15] and *BigQuery GHTorrent* [16] data sets. For our quantitative analyses, we mainly used data from the BigQuery GH data set, retrieved July 20, 2016 and February 9, 2017. To estimate how often developers use snippets from SO without attribution (RQ2), we followed two different approaches. First, we used a token-based code clone detector, the *PMD Copy-Paste Detector* (CPD, version

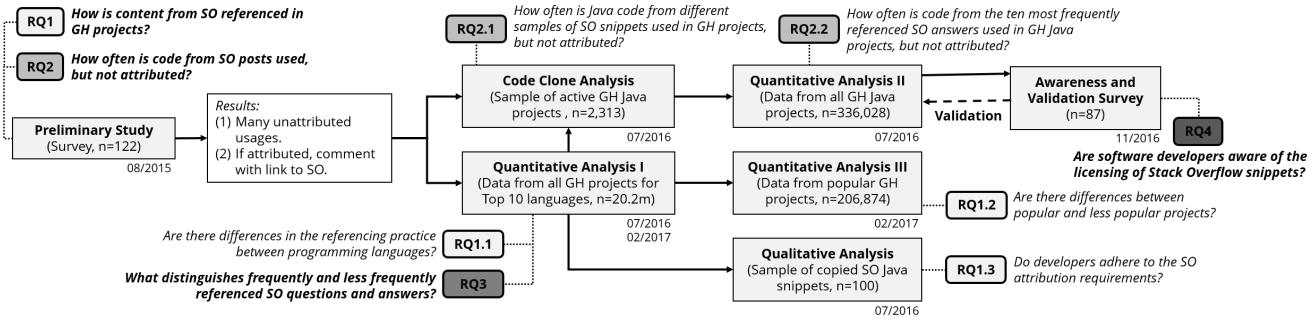

Fig. 1. High-level research design to study the usage and attribution of Stack Overflow (SO) code snippets in GitHub (GH) projects.

5.4.1) [17], to find unreferenced usages of three different sets of SO code snippets in a random sample of popular GH Java projects. Second, we created regular expressions matching the code snippets of the ten most frequently referenced Java answers on SO. With the help of BigQuery, we utilized these regular expressions to find unreferenced usages in all Java projects in the data set.

III. Summarized Results

In this section, we summarize the results of each study. We provide the raw data and all analysis scripts as supplementary material [18].

1) Preliminary Study (RQ1+2): The goal of the preliminary study was to get first insights into developers' practices regarding the usage and attribution of code from SO. We contacted 1,000 randomly selected users who were active on both SO and GH and received 122 responses (12.2% response rate). Participants reported that the last time they copied or adapted a code snippet from SO, half of them (49%) did not attribute its origin; 40% added a source code comment with a link to the corresponding question or answer.

2) Quantitative Analysis I+III (RQ1+3): Using BigQuery and a regular expression, we searched all source code files of different programming languages in non-fork GH repositories for references to SO. There were on average twice as many references to whole SO threads than to specific answers. On average, 3.22% of all repositories and 7.33% of all popular repositories (more than 21 watchers, 99% quantile for all languages) contained a reference to SO. R, Python, C#, and Objective-C files contained more references to SO compared to the other analyzed languages. Frequently referenced questions and answers had a significantly higher view count and score ($p_w<0.001$, $|d|\geq 0.5$); frequently referenced answers had significantly more code blocks, but the effect was only small.

3) Code Clone Analysis (RQ2.1): We used CPD to find unreferenced usages of three sets of SO code snippets in a sample of popular GH Java projects (more than 29 watchers, 99% quantile for Java). We found that in our sample of Java projects (n=2,313), 207 repositories (9%) contained a copy of a snippet from one of the three SO snippet sets. Only 23% of the matched files contained a reference to SO.

4) Quantitative Analysis II (RQ2.2): To complement the above results, we applied a second approach for finding unreferenced usages of SO code snippets using regular expressions and the BigQuery GH data set. We searched for copies of the snippets from the ten most frequently referenced SO Java answers in all non-fork Java GH projects. At most 27% of the identified usages were attributed.

5) Qualitative Analysis (RQ1.3): We manually analyzed a random sample of SO references in Java files to assess how developers refer to SO content in source code comments (n=100). Most comments included only a link to the corresponding answer without naming the author of the code. Only 11 out of 97 analyzed comments explicitly named SO as the source.

6) Awareness and Validation Survey (RQ4): To complement the quantitative results, we conducted a second online survey to investigate the awareness of developers regarding the licensing of SO content. We further used this survey to detect false positives in our analysis for RQ2.2. We contacted 739 owners of GH repos in which we found matches of SO snippets (87 responses, 11.8%). Most developers (75%) were not aware of the licensing of code published on SO; only 32% of the participants were aware of the attribution requirement. Most participants answered that the code has either been copied from SO, or they did not remember.

IV. Conclusion and Future Work

In our analysis of unattributed usages of SO code snippets, we always chose the most conservative estimates. Using CPD, we found that only 23% of the identified clones of Java snippets included a reference to SO. Using BigQuery and regular expressions for the ten most frequently referenced Java snippets, our estimate was 27% attributed usages. Thus, we think that one third is a reasonable upper bound for the amount of attributed usages. We further investigated how often SO URLs are present in source code files of different programming languages. On average, 3.22% of all analyzed repositories and 7.33% of the popular ones contained a file with a reference to SO. Depending on the project's license, this may lead to legal issues for the projects. Our second survey has shown that many developers admit copying code from SO without attribution, but they are not aware of the licensing and its implications. The next steps of our research are to automate and scale the extraction of copyable snippets form SO and the detection of unattributed usages in GH projects. This 'reverse engineering' of the missing link to SO can help developers mitigating possible maintenance and legal issues as described in the introduction.

REFERENCES

[1] Stack Exchange Inc., "Stack Exchange data explorer," http://data.stackexchange.com/stackoverflow/query/new, last accessed: August 23, 2016.

[2] D. Yang, A. Hussain, and C. V. Lopes, "From query to usable code: An analysis of Stack Overflow code snippets," in *13th International Workshop on Mining Software Repositories (MSR'16)*. ACM, 2016, pp. 391–402.

[3] J. Brandt, M. Dontcheva, M. Weskamp, and S. R. Klemmer, "Example-centric programming: Integrating web search into the development environment," in *SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, 2010, pp. 513–522.

[4] Stack Exchange Community Wiki, "Do I have to worry about copyright issues for code posted on Stack Overflow?" http://meta.stackexchange.com/q/12527, December 2013.

[5] S. Ruport, "Can we get some explicit clarification on the *intended* legal usage of code from SO answers?" http://meta.stackoverflow.com/q/286582, February 2015.

[6] csharptest.net, "What is up with the source code license on Stack Overflow?" http://meta.stackexchange.com/q/25956, October 2009.

[7] samthebrand, "The MIT license – clarity on using code on Stack Overflow and Stack Exchange," http://meta.stackexchange.com/q/271080, December 2015.

[8] ——, "A new code license: The MIT, this time with attribution required," http://meta.stackexchange.com/q/272956, January 2016.

[9] C. Vendome, "A large scale study of license usage on GitHub," in *37th International Conference on Software Engineering (ICSE'15)*, 2015.

[10] C. C. Corporation, "Compatible licenses," https://creativecommons.org/share-your-work/licensing-considerations/compatible-licenses/, last accessed: February 7, 2017.

[11] GitHub Inc., "GitHub Developer – API," https://developer.github.com/v3/, last accessed: August 9, 2016.

[12] Stack Exchange Inc., "Stack Exchange API v2.2," https://api.stackexchange.com/docs, last accessed: August 9, 2016.

[13] G. Gousios, "The GHTorent dataset and tool suite," in *10th Working Conference on Mining Software Repositories (MSR'13)*, 2013.

[14] Stack Exchange Inc., "Stack Exchange Data Dump January 1, 2016," https://archive.org/details/stackexchange/.

[15] Google CloudPlatform, "GitHub data," https://cloud.google.com/bigquery/public-data/github, last accessed: August 9, 2016.

[16] G. Gousios, "GHTorrent on the Google cloud," http://ghtorrent.org/gcloud.html, last accessed: February 7, 2017.

[17] PMD source code analyzer, "Finding duplicated code," http://pmd.github.io/pmd-5.5.1/usage/cpd-usage.html.

[18] S. Baltes, R. Kiefer, and S. Diehl, "Stack Overflow code snippets in GitHub projects – supplementary material," http://snippets.sbaltes.com.