

Empirical Study to Track Code Flow from StackOverflow To Github

Kulendra Kumar Kaushal*, Rutwik Kulkarni†

* Department of Computer Science

Virginia Tech, Blacksburg, Virginia Tech 24060

Email: kulendra@vt.edu

†Department of Computer Science

Virginia Tech, Blacksburg, Virginia Tech 24060

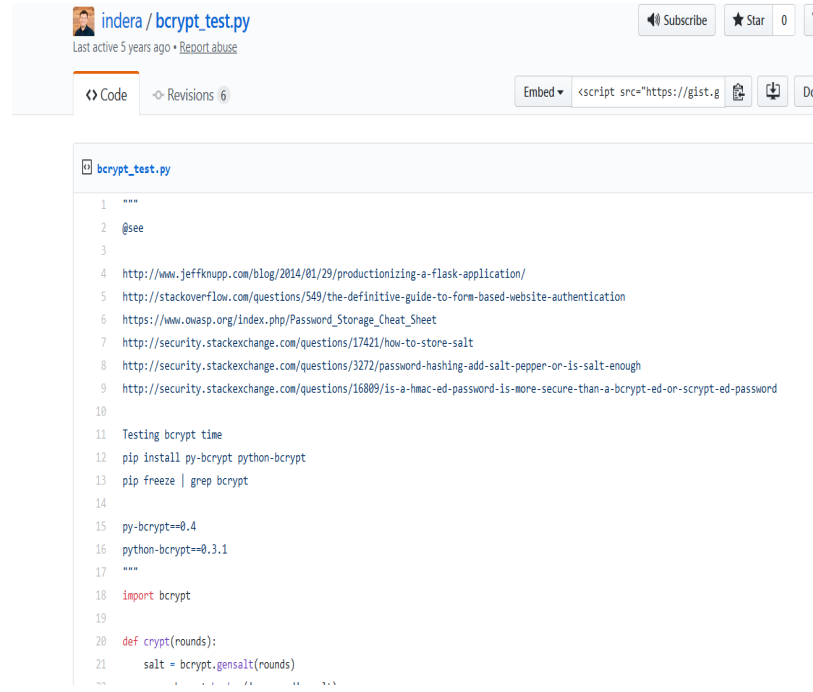
Email: rutwikk@vt.edu

Abstract—Today, most of the software developers resolve their programming queries by posting questions on StackOverflow. Many popular software libraries are open-source and have their code published in repositories on Github. Preliminary observation showed us that most of the developers cite StackOverflow questions in their Github repository whenever they use it to solve their problems. In this project, we conduct an empirical study to investigate the presence of StackOverflow snippets in the source code of popular public Github repositories. In order to conduct the study, we combine StackOverflow and Github publically available data dump, perform exploratory analysis on the extracted data and analyze the extent to which the code was copied from one website to another. We also investigate whether the most upvoted answer of a StackOverflow question is the most referenced answer.

I. INTRODUCTION

StackOverflow has become a prime platform for the developers to resolve their queries and has posts on nearly every topic in the computer science domain. Github is a popular version control software that hosts the repositories of open source projects. It has been generally observed that whenever a developer refers to the code on StackOverflow, he cites the referred question as a comment in his code published in the Github repository. Figure 1 shows the StackOverflow question “The definitive guide to form-based website authentication” cited by the developer of Indera repository on Github. We also hypothesize that certain questions on Stackoverflow have been answered by using the code snippets available on Github. In this study, we trace the flow of code from Github to StackOverflow and vice versa. We conduct this study by answering the following research questions.

- RQ1: What are the StackOverflow answers most frequently referenced by github codebases?
- RQ2: Is the most upvoted answer to any StackOverflow questions, the most referenced answer?
- RQ3: What is the extent to which the code is copied from StackOverflow to Github?
- RQ4: Are the answers to RQ1, RQ2, and RQ3 consistent for different programming languages?



```
1  """
2  @see
3
4  http://www.jeffknupp.com/blog/2014/01/29/productionizing-a-flask-application/
5  http://stackoverflow.com/questions/549/the-definitive-guide-to-form-based-website-authentication
6  https://www.owasp.org/index.php/Password_Storage_Cheat_Sheet
7  http://security.stackexchange.com/questions/17421/how-to-store-salt
8  http://security.stackexchange.com/questions/3272/password-hashing-add-salt-pepper-or-is-salt-enough
9  http://security.stackexchange.com/questions/16809/is-a-hmac-ed-password-is-more-secure-than-a-bcrypt-ed-or-scrypt-ed-password
10
11  Testing bcrypt time
12  pip install py-bcrypt python-bcrypt
13  pip freeze | grep bcrypt
14
15  py-bcrypt==0.4
16  python-bcrypt==0.3.1
17  """
18  import bcrypt
19
20  def crypt(rounds):
21      salt = bcrypt.gensalt(rounds)
22      ...
```

Fig. 1. Github repository citing StackOverflow question

A. What are the StackOverflow answers most frequently referenced by Github codebases?

Different users provide different answers to a question on StackOverflow. StackOverflow also allows the users to upvote an answer if they like it. Preliminary observation showed us that some of the popular Github repositories used a less upvoted answer as a solution to their problem instead of the most popular one. Addressing this research question will help us build a system that will show the user the most referenced answer along with the most upvoted one.

B. Is the most upvoted answer to any StackOverflow questions, the most referenced answer?

Every StackOverflow question has multiple answers that are posted by various developers. It is a common assumption that

the most upvoted answer is used or referenced by maximum users. By analyzing the most referred answers to the questions we validate the assumption that whether the most upvoted answer is the most referred answer by the popular GitHub repositories

C. What is the extent to which the code is copied from StackOverflow to Github?

There have been numerous instances where code snippets from StackOverflow have been used in Github projects[2]. Further, there has always been a debate of upto what extent developers copy answers directly from different StackOverflow questions. Popular Github repositories use the StackOverflow answers for the following reasons.

- The developers copy the code entirely as a solution to their problem.
- The developers follow a similar approach as that of the referenced answer.
- The developer cites the answer because of it a solution to the current/potential bug We aggregate the answers in the groups given in the list above and perform exploratory data analysis on the same

D. Cross-language analysis of the above research questions

Analysis of the RQ1, RQ2, and RQ3 across different programming languages has not been conducted until now. Currently, we have combined the data for StackOverflow Questions and Github repositories for Java and JavaScript language.

II. BACKGROUND AND RELATED WORK

In the development life cycle, the reuse of software components is always recommended. It saves time and also minimizes the effort. Therefore, the developer community tries to avoid reinventing the wheel. Based on the concept of reusability, communities try to build libraries and modules which can be reused. Further, the documentation of many projects is often not very clear and also becomes obsolete with time [4]. The social media platforms play a critical role in helping developers. These platforms help the developers to solve their programming queries that rise because of obsolete documentation [11]. StackOverflow has become a popular Q&A forum being actively used by millions of developers [14]. StackOverflow helps developers in implementing and learning API documentation related to Android, Java, and GWT [14] [10]. Researchers have conducted various studies in order to find top contributors on StackOverflow [9]. They have also tried to track the use of StackOverflow in academic papers [8]. Open-source software development has been proven to be a revolution in the way knowledge has spread among the developer community [6]. The work done by the authors of [3] shows how fellow developers learn from popular open-source developers. It also highlights the awareness of constantly being watched by peers and the way in which it affects the pattern of contribution by developers. All these studies have highlighted the importance of StackOverflow and Github in the software

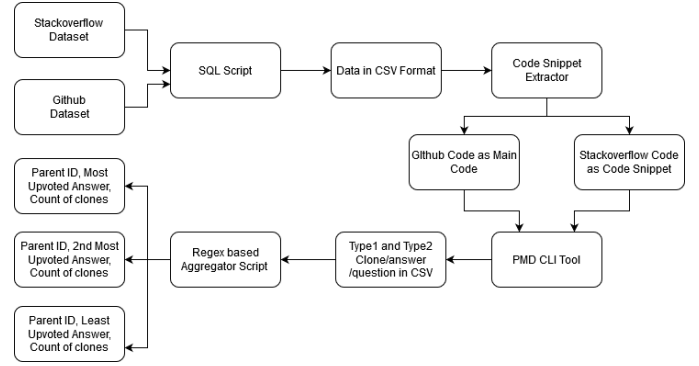


Fig. 2. System Architecture

community. The work done by the authors of [13] is closest to our research direction. It studies the correlation between developer's activity on StackOverflow and the corresponding commit history of the same developer on Github. However, it is focussed on the profile of the same StackOverflow and Github developers. Stack overflow questions often have multiple answers for any given question. The answer which is most referred to by the developed is still undiscovered. We can observe that none of the researchers have studied the reason for the citation of StackOverflow answers on Github. We can also see that most of the researchers have concentrated on the issues in Java programming language for their analysis. Therefore, the above-mentioned research questions still remain open and can be explored as a course project.

III. IMPLEMENTATION

A. System Architecture

In this section, we discuss the architecture of our system. Figure 2 describes the architecture of our system. Many public datasets containing information about StackOverflow posts and Github repositories are available. We used a SQL query to obtain the StackOverflow questions and their corresponding referenced Github repositories. The result obtained from the SQL query is stored in the CSV file. Any StackOverflow answer consists of a brief explanation along with the code snippet supporting it. We extract code snippets from both StackOverflow answers and the referenced GitHub repositories for further code clone detection. Code clone detection is done using the PMD [1] code clone detection tool. The results of clone detection are passed downstream to the aggregator and analysis scripts for further inference and evaluation. The consequent sections will explain each implemented step in detail.

B. Database Curation

Multiple open data corpus of StackOverflow dump and Github extracts are available on different online platforms like Kaggle, Stack Exchange, Google Bigquery, etc. Each of the platforms has its own mechanism of providing data. Stack Exchange provides data via Stack Exchange Data Explorer (SEDE) [5]. SEDE can only extract only 50,000 rows at once.

Additionally, SEDE doesn't provide any native API to extract data. It also protects itself from abuse with CAPTCHAs. Therefore, it can not be queried using a web request. Data dump present on Kaggle can be accessed via CSV files or Kernels. Google BigQuery [12] data dump can be queried using a SQL query. Additionally, Google allows users to create tables and views on top of its data dump which can be used for optimizing the query. We decided to use Google BigQuery to extract data because different open-source curated datasets are present at a single place and can be combined without any additional overhead. The StackOverflow data and Github data are combined based on the unique question ID of StackOverflow posts. Whenever a developer references any StackOverflow post, he puts the link of that particular StackOverflow post as a comment in the code repository. We filter Github repositories having StackOverflow URLs. We extract question ID from the files of those Github repositories using a regex expression. Finally, we join the StackOverflow data with corresponding Github repositories using the question ID column. Data belonging to different programming languages are present in different tables. Google imposes an upper bound on resource allocation for a single query and also on the number of free queries that a user can send to BigQuery. We need to store the results using table views because tables have a very large number of rows and an unoptimized query can exhaust the freely available resources. Likewise, the 1st, 2nd, and 3rd most upvoted answers for a StackOverflow question can be extracted only using joins instead of a correlational subquery.

```
#standardSQL
SELECT      a.id ,
            title ,
            body ,
            content ,
            parent_id ,
            favorite_count favs ,
            view_count      views ,
            score ,
            accepted_answer_id ,
            post_type_id ,
            sample_repo_name ,
            sample_path

FROM
    `bigquery-PUBLIC-data .
stackoverflow . posts_answers ` a
INNER JOIN
    (
        SELECT cast(
            regexp_extract
            (content ,
            r`stackoverflow . com/ questions / ([0 - 9] +) / `
            ) AS int64 ) id ,
            sample_repo_name ,
            sample_path ,
            content
```

```
FROM      `fh-bigquery . github_extracts . contents_java `
WHERE     content LIKE `%stackoverflow . com/ questions / %`
ON        a . parent_id = b . id
WHERE     parent_id IS NOT NULL
AND
    (
        SELECT count (*)
        FROM
            `bigquery-PUBLIC-data . stackoverflow . posts_answers `
        WHERE
            a . parent_id = t2 . parent_id
            AND
            t2 . score >= a . score ) <= 3
ORDER BY  parent_id ,
            score limit 10000

Listing 1. Query for data extraction
```

Listing 1 shows the SQL query used for extraction of data from publicly available datasets

The extracted data for Java language has 10,000 rows containing StackOverflow answers and their corresponding Github repositories. The data contains 1200 unique questions. The total number of unique answers in this extracted dataset is 3552. Further analysis of the data reveals to us that there are 43 questions in the dataset with only one answer, while 162 questions have two unique answers. It can be seen that 995 out of 1200 questions have 3 unique answers. Thus, we can find the most referenced and most favorite answer form this dataset. The data for javascript language has 1000 rows with 212 unique questions and 227 unique answers.

C. Code Clone Detection

The data consists of the code snippets from StackOverflow answers and the content of files in which the question of these answers is referenced. For checking whether the clone of an answer is present in the GitHub repositories we used PMD code clone detection tool. PMD calculates the presence of clones in two files using a token-based approach. We arrange the code snippets of the answers and the GitHub codebase into separate files. Each file having a name like “key_main_code” contains the content from Github repositories. Similarly, a file having a name in the format “key_code_snippet” contains the code snippets for StackOverflow answers. The key for StackOverflow answer and its corresponding Github file is the same. PMD calculates clones among all the files in a folder and the result is saved in a CSV file. In order to extract the relevant clones (clones from StackOverflow to Github) we use the unique key discussed above. PMD supports type 1 and type 2 clone detection. In type 1 clone detection an exact match is found while in type 2 the values of literals and the names of identifiers are ignored. To assess our RQ3, we calculate type 1 and type 2 clones for Java. PMD can find only type 1 clones for JavaScript.

D. Benchmark for Cross-language Analysis

PMD follows a token-based approach for code clone detection. To detect meaningful clones from the data it is necessary

Program	Java	JavaScript	Ratio
BST	502	146	3.43
Hello World-Text	35	8	4.375
Hello World-Server	154	70	2.2
Hello World Graphical	417	8	52.125
Swap	74	39	1.8974
While Loop	47	34	1.382
For loop	77	89	0.8651
MatMul	277	338	0.819
Object Creation	26	44	0.59

TABLE I

TOKEN COUNT FOR DIFFERENT PROGRAMS IN JAVA AND JAVASCRIPT

to decide the minimum number of same tokens detected in two files. A semantically same code can be written using a different number of tokens in different languages. Therefore, benchmarking the minimum number of the same tokens so that two snippets can be labeled as clones across multiple languages is necessary. Initially we perform manual analysis to find the number of clones for different numbers of tokens. We found that a JavaScript clone with more than 10 tokens was meaningful. As manual analysis is prone to errors, we perform analysis with the help of code snippets in Rosetta Github repository [7]. The Rosetta code repository contains snippets of the same code written in various programming languages. We analyzed the simple code for JavaScript and Java across various domains such as web development, graphical applications, simple scripting, and basic array operations. Stackoverflow contains questions and answers from different domains, therefore analyzing the snippets across multiple domains helps us perform accurate benchmarking. From the table III-D, we can see the median of the ratio of the number of tokens in Java to JavaScript is 1.89. Combining the results from manual analysis and keeping the baseline tokens for JavaScript as 10, we obtain baseline tokens for Java as 19 (1.89×10).

IV. EVALUATION

The data obtained from the clone detection is used to answer our research questions as follows

A. RQ1:

From the code clone detection section, we obtain the count of clones for every answer in the StackOverflow database. After obtaining the count, we aggregate the answers by their respective question. Thus, we were able to identify an answer to a specific question that has maximum references in popular Github repositories. The data of questions and their respective favorite answer is stored in CSV format.

B. RQ2:

Based on the count of clones for StackOverflow answers, grouped by their respective questions we perform analysis of RQ2. The most favorite answer to any question is the one with maximum upvotes. This answer can be extracted from the 'Score' parameter in the dataset. The most referenced

answer is found from RQ1. Considering type 1 clones for Java references, we find the following references from our data.

- 1) 66 questions had their most upvoted answer referenced once or more in popular Github repositories.
- 2) 56%(37/66) of times the most upvoted answer has the highest number of referenced Github repositories while 31% (20/66) of times, 2nd most upvoted answer has a higher number of referenced Github repositories for the same question.
- 3) 13% (9/66) questions the least favorite Answer has the highest number of referenced Github repositories as compared to other answers to the same question.

Considering type 2 clones for Java references, we find the following references from our data.

- 1) 47 Questions had some or the other answer referenced in Github repositories.
- 2) The most favorite Answer was referenced 63.8% (30/66) number of times.
- 3) 36.17% (17/47) of questions the most referenced answer was not the most favorite one
- 4) About 19% (9/47) questions the least favorite Answer is the most referenced

For the snippets and Github repositories in JavaScript, we found that 17 questions had their most upvoted answer referenced once or more in popular Github repositories

C. RQ3:

To analyze the extent to which the code is copied from Stack Overflow to Github In section ??, we mention the reasons for which StackOverflow questions are referenced in the Github repositories. 1200 Java questions in our dataset had references in the Github repositories. We were not able to find the type 2 clones for a baseline of 19 tokens due to limited computing resources. Therefore, we set a baseline tokens of 30 for analysis. We found 20 questions having their answer cloned as type 1 clones and 47 questions having their answer as type 2 cloned from our analysis. Thus, we can say that for Java the reference for a question is not necessarily because a code snippet from Stackoverflow answer is cloned in the repository.

D. RQ4:

To perform the cross-language analysis, we perform baseline token analysis as discussed in section III-D. We found that approximately one token in JavaScript language is equivalent to two tokens in Java. Additionally, we observe that for Java, 5.5% (66/1200) of the questions have their most upvoted answer referenced once or more while in the case of JavaScript, the percentage is 8.5% (17/212). From Rosetta equivalent token analysis for Java and JavaScript as a baseline, we can say that for JavaScript most upvoted answers are referenced more as compared to Java.

V. CONCLUSIONS AND FUTURE WORK

In this empirical study, we were successfully able to find the most referenced answer to any StackOverflow question.

Along with the most favorite answer we can also show the most referenced answer to the user. We also found that it is not necessary for the most favorite answer to be most referenced. Thus, the most referenced answer can have more advantages than the most favorite one. As it is used by popular repositories, it can provide a more optimized solution to the problem. To identify whether the trend of most referenced answers not being the most favorite is not limited to a single programming language we also performed the cross-language analysis of our research questions. We would further obtain the relationship between the most referenced answer and accepted answer and also extend this study for multiple languages. Thus, our research adds another dimension of the most referenced answer that can guide developers in choosing the most appropriate answer for solving their problem.

REFERENCES

- [1] ARCELLI FONTANA, F., ZANONI, M., RANCHETTI, A., AND RANCHETTI, D. Software clone detection and refactoring. *ISRN Software Engineering 2013* (2013).
- [2] BALTES, S., AND DIEHL, S. Usage and attribution of stack overflow code snippets in github projects. *Empirical Software Engineering 24*, 3 (2019), 1259–1295.
- [3] DABBISH, L., STUART, C., TSAY, J., AND HERBSLEB, J. Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (2012), pp. 1277–1286.
- [4] LETHBRIDGE, T. C., SINGER, J., AND FORWARD, A. How software engineers use documentation: The state of the practice. *IEEE software 20*, 6 (2003), 35–39.
- [5] LORD, L., SELL, J., BAGIROV, F., AND NEWMAN, M. Survival analysis within stack overflow: Python and r. In *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)* (2018), IEEE, pp. 51–59.
- [6] MOCKUS, A., FIELDING, R. T., AND HERBSLEB, J. A case study of open source software development: the apache server. In *Proceedings of the 22nd international conference on Software engineering* (2000), pp. 263–272.
- [7] NANZ, S., AND FURIA, C. A. A comparative study of programming languages in rosetta code. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering* (2015), vol. 1, IEEE, pp. 778–788.
- [8] OSBORNE, M. J. Academic papers using stack overflow data.
- [9] OSBORNE, M. J. Top users on stackoverflow: slackers or superstars?
- [10] PARNIN, C., TREUDE, C., GRAMMEL, L., AND STOREY, M.-A. Crowd documentation: Exploring the coverage and the dynamics of api discussions on stack overflow. *Georgia Institute of Technology, Tech. Rep 11* (2012).
- [11] STOREY, M.-A., TREUDE, C., VAN DEURSEN, A., AND CHENG, L.-T. The impact of social media on software engineering practices and tools. In *Proceedings of the FSE/SDP workshop on Future of software engineering research* (2010), pp. 359–364.
- [12] TIGANI, J., AND NAIDU, S. *Google BigQuery Analytics*. John Wiley & Sons, 2014.
- [13] VASILESCU, B., FILKOV, V., AND SEREBRENIK, A. Stackoverflow and github: Associations between software development and crowdsourced knowledge. In *2013 International Conference on Social Computing* (2013), IEEE, pp. 188–195.
- [14] WU, Y., WANG, S., BEZEMER, C.-P., AND INOUE, K. How do developers utilize source code from stack overflow? *Empirical Software Engineering 24*, 2 (2019), 637–673.