SOTorrent: Studying the Origin, Evolution, and Usage of Stack Overflow Code Snippets

Sebastian Baltes
University of Trier, Germany
research@sbaltes.com

Christoph Treude
University of Adelaide, Australia
christoph.treude@adelaide.edu.au

Stephan Diehl
University of Trier, Germany
diehl@uni-trier.de

Abstract—**Stack Overflow (SO) is the most popular question-and-answer website for software developers, providing a large amount of copyable code snippets. Like other software artifacts, code on SO evolves over time, for example when bugs are fixed or APIs are updated to the most recent version. To be able to analyze how code and the surrounding text on SO evolves, we built *SOTorrent*, an open dataset based on the official SO data dump. *SOTorrent* provides access to the version history of SO content at the level of whole posts and individual text and code blocks. It connects code snippets from SO posts to other platforms by aggregating URLs from surrounding text blocks and comments, and by collecting references from GitHub files to SO posts. Our vision is that researchers will use *SOTorrent* to investigate and understand the evolution and maintenance of code on SO and its relation to other platforms such as GitHub.**

I. INTRODUCTION

Stack Overflow (SO) is the most popular question-and-answer website for software developers. Many of its over 40 million posts [1] contain code snippets together with explanations [2]. Those snippets do not exist in isolation but are actively reused by developers in their software projects, regardless of maintainability, security, and licensing implications [3]–[10]. Yet, there is still a lack of knowledge on how exactly SO code snippets are sourced from and reused on other platforms. Understanding the evolution of SO content, together with its origin and usage in software projects, is crucial to identify outdated information, detect compatibility issues, and prevent copy-and-paste bugs.

Similar to other software artifacts such as source code files and documentation [11]–[14], text and code on SO evolve over time, e.g., when the SO community fixes bugs in code snippets or updates documentation to match new API versions. Since the inception of SO in 2008, a total of 14.8 million SO posts have been edited after their creation—21,601 of them more than ten times. While many SO posts contain code, the evolution of code snippets on SO differs from the evolution of entire software projects. Most snippets are relatively short [15] and many of them cannot compile without modification [2]. In addition, SO does not provide a version control or bug tracking system for code snippets, forcing users to rely on the commenting function or additional answers to voice concerns about a snippet.

II. SOTORRENT: MSR MINING CHALLENGE 2019

SOTorrent is an open dataset based on data from the official SO data dump [1] and the Google BigQuery GitHub (GH) dataset [16] that enables researchers to analyze the version history of SO posts at the level of individual text and code blocks (see Figure 1 for exemplary posts). The official SO data dump [1] keeps track of different versions of entire posts, but does not contain information about differences between versions at a more fine-grained level. In particular, extracting different versions of the same code snippet from the history of a post is challenging and required us to develop a complex strategy, involving the evaluation of 134 different string similarity metrics [15]. Beside providing access to the version history, our dataset links SO posts to external resources in two ways: (1) by extracting linked URLs from text blocks of SO posts and from post comments and (2) by providing a table with links to SO posts found in the source code of open source GH projects. This table can be used to connect *SOTorrent* to GH datasets such as *GHTorrent* [17]. Analyses can be based on *SOTorrent* alone or expanded to include data from other resources (see Figure 2). Questions that are, to the best of our knowledge, not sufficiently answered yet include:

- How are code snippets on SO maintained?
- How many clones of code snippets exist inside SO?
- How can we detect buggy versions of SO code snippets and find them in GH projects?
- How frequently are code snippets copied from external sources into SO and then co-evolve there?
- How do snippets copied from SO to GH co-evolve?
- Does the evolution of SO code snippets follow patterns?
- Do these patterns differ between prog. languages?
- Are the licenses of external sources compatible with SO's license (CC BY-SA 3.0)?
- How many code blocks on SO do not contain source code (and are only used for markup)?
- Can we reliably predict bug-fixing edits to code on SO?
- Can we predict popularity of SO code snippets on GH?

SOTorrent is available on Zenodo as a CSV database dump [18] together with instructions on how to import the dataset. Moreover, the dataset is available as a Google Big-Query dataset [19], which allows to execute complex queries without the need to import the dataset (1 TB of queries per month are free). We also published the source code of the software that we used to build [20], [21] and analyze [22], [23] *SOTorrent*. More information about the dataset can be found in the corresponding research paper [15].
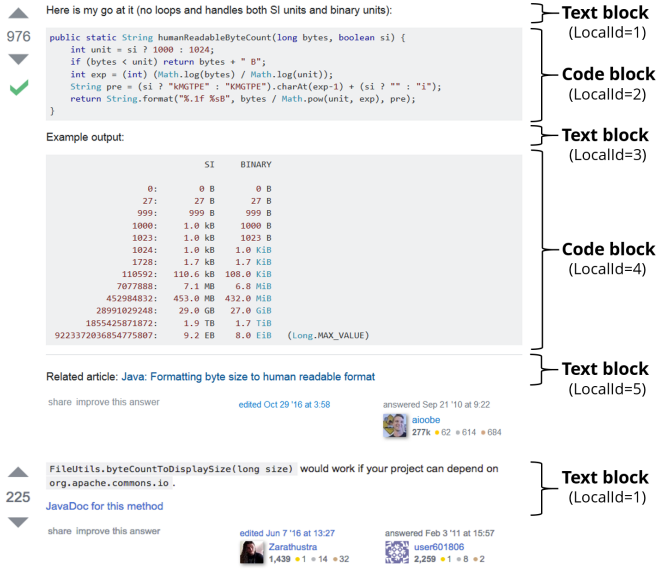
APPENDIX A
FIGURES

Fig. 1. Exemplary Stack Overflow answers with code blocks (top, 3758880) and with inline code (bottom, 4888400). The `LocalId` represents the position in the post.
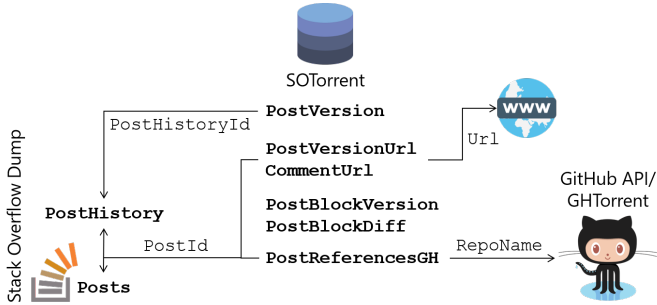

Fig. 2. Connection of *SOTorrent* tables to other resources.

APPENDIX B
DATA COLLECTION AND DATABASE SCHEMA

The *SOTorrent* dataset contains all tables from the official Stack Overflow data dump. However, that dump does only provide the version history at the level of whole posts as Markdown-formatted text. To analyze how individual text or code blocks evolve, we needed to extract individual blocks from that content. This extraction also enabled us to collect links to external sources from the identified text blocks.

In the SO dump, one version of a post corresponds to one row in the table `PostHistory`. However, that table does not only document changes to the content of a post, but also changes to metadata such as tags or events such as closing of posts. Since our goal was to analyze the evolution of SO posts at the level of whole posts and individual post blocks, we had to filter and process the available data. First,

we selected edits that changed the content of a SO post, identified by their `PostHistoryTypeId` [24] (2: *Initial Body*, 5: *Edit Body*, 8: *Rollback Body*). We linked each filtered version to its predecessor and successor and stored it in table `PostVersion`.

The content of a post version is available as Markdown-formatted text. We split the content of each version into text and code blocks and extracted the URLs from all text blocks using a regular expression (table `PostVersionUrl`). We also extracted the URLs from all comments in the SO data dump (table `CommentUrl`). Beside the extracted URLs, those tables provide information about the link type (e.g., bare, Markdown, or HTML), link position (top, middle, or end of post/comment), and certain URL components such as the root domain, query string, or fragment identifier (if present). To reconstruct the version history of individual post blocks (table `PostBlockVersion`), we established a linear predecessor relationship between the post block versions using a string similarity metric that we selected after a thorough evaluation [15]. For each post block version, we computed the line-based difference to its predecessor, which is available in table `PostBlockDiff`.

We also extracted the version history of question titles from table `PostHistory`. Table `TitleVersion` links all title versions to their predecessors and successors and further provides the corresponding Levenshtein distances (columns `PredEditDistance` and `SuccEditDistance`).

One row in table `PostReferenceGH` represents one link from a file in a public GH repository to a post on SO. To extract those references, we utilized Google BigQuery, which allows to execute SQL queries on various public datasets, including a dataset with all files in the default branch of GH projects [16]. To find references to SO, we again applied a regular expression and mapped all extracted URLs to their corresponding sharing link (ending with `/q/<id>` for questions and `/a/<id>` for answers), storing that link together with information about the file and the repository in which the link was found in table `PostReferenceGH`. We ignored other links referring to, e.g., users or tags on SO.

Version 2018-08-28 of the *SOTorrent* dataset contains the version history of all 40,606,950 questions and answers in the official SO data dump published June 5, 2018 [25]. It contains 63,914,798 post versions, 122,673,430 text block versions, and 77,578,494 code block versions, ranging from the creation of the first post on July 31, 2008 until the last edit on June 3, 2018. We extracted links to 11,775,659 distinct URLs from 20,518,181 different post block versions and 4,104,869 distinct URLs from 6,856,777 different comments. Moreover, we identified 6,035,737 links to SO posts in 436,615 public GH repositories.

Our project website [1] lists all dataset versions and contains more information on the database layout, including the complete database schema.

---

[1] http://sotorrent.org

To illustrate how researchers can use *SOTorrent* to analyze the evolution of SO posts, we investigate one of the most popular Java answers on SO, which is depicted in Figure 1. It is the accepted answer for the question *"How to convert byte size into human readable format in java?"*.[2] The following SQL queries are tested on a MySQL 5.7 database system with *SOTorrent* 2018-08-28, but will also work with later versions. Since some *SOTorrent*-specific IDs (e.g., `PostVersionId` or `PostBlockVersionId`) may change between dataset versions, it is recommended to use the IDs from the official Stack Overflow data dump when possible (e.g., `PostId` or `PostHistoryId`), which are included as foreign keys and are stable across all dataset versions. Exemplary BigQuery queries can be found in an earlier blog post about *SOTorrent*[3]. We start our investigation by retrieving all post block versions of the above-mentioned answer using its `PostId`:

```sql
SELECT PostHistoryId, PostBlockTypeId, LocalId, ...
    # see Fig. 3 for all selected columns
FROM PostBlockVersion
WHERE PostId=3758880
ORDER BY PostHistoryId ASC, LocalId ASC;
```

Part (1) of Figure 3 shows the result of the above query for the two most recent post versions. In our *SOTorrent* paper [15], we defined post block lifespans as chains of connected post block versions that are predecessors of each other. Those chains can be easily retrieved from the database, because each post block version points to its `RootPostHistoryId` and `RootLocalId`. Those columns uniquely identify the first post block version in the chain. For part (2) of the figure, we only selected versions of the code snippet in the answer in which the content was actually modified (not all post blocks are modified in all post versions):

```sql
SELECT Id, PostHistoryId, LocalId, Content, Length, ...
    # see Fig. 3 for all selected columns
FROM PostBlockVersion
WHERE RootPostHistoryId=7873162 AND RootLocalId=2
  AND (PredEqual IS NULL OR PredEqual = 0)
ORDER BY PostHistoryId ASC;
```

To further see which lines of a code snippet were changed in the last edit, we can utilize table `PostBlockDiff`. The result of the following query is shown in part (3) of the figure:

```sql
SELECT PostHistoryId, LocalId, PostBlockDiffOperationId,
    Text
FROM PostBlockDiff
WHERE PostHistoryId=7875126 AND LocalId=2;
```

We can also use *SOTorrent* to retrieve files on GitHub that reference this particular Stack Overflow post:

```sql
SELECT RepoName, Branch, Path, FileExt, Copies, PostId,
    SOUrl, GHUrl
FROM PostReferenceGH
WHERE PostId=3758880;
```

The result of this query is shown in Figure 4. To retrieve links from all text block versions of the post, we can use table `PostVersionUrl`:

```sql
SELECT *
FROM PostVersionUrl
WHERE PostId=3758880;
```

In this case, only one post block version contains a link, which refers to a blog post with the same snippet (see Figure 1). To retrieve the title versions of the question that started the thread, we can use the following query:

```sql
SELECT *
FROM TitleVersion
WHERE PostId=(
  SELECT ParentId
  FROM Posts
  WHERE Id=3758880
);
```

For this thread, however, the title has never been changed. Retrieving all links from comments to this particular post is as simple as:

```sql
SELECT *
FROM CommentUrl
WHERE PostId=3758880;
```

This query reveals one link in a comment, pointing to a class in the iOS API providing a functionality similar to the snippet in the post.

In our previous *SOTorrent* paper [15], we described a close relationship between post edits and comments. To support a further investigation of this relationship, we wrote a blog post[4] showing how to create a new table `EditHistory`, which aggregates all title and body edits of Stack Overflow posts, together with post comments. Using this table (and a helper table `Threads`), one can easily retrieve the edit and comment history of individual threads (see blog post for more details):

```sql
SELECT * FROM EditHistory
WHERE PostId IN (
  SELECT PostId FROM Threads WHERE ParentId = (
    SELECT ParentID FROM Threads
    WHERE PostId=3758880
    # the question PostId 3758606 yields the same result
)) ORDER BY CreationDate;
```

REFERENCES

[1] Stack Exchange Inc, "Stack Exchange Data Dump 2018-06-05," 2018. [Online]. Available: https://archive.org/details/stackexchange/

[2] D. Yang, A. Hussain, and C. V. Lopes, "From Query to Usable Code: An Analysis of Stack Overflow Code Snippets," in *13th International Conference on Mining Software Repositories (MSR 2016)*, M. Kim, R. Robbes, and C. Bird, Eds. Austin, TX, USA: ACM, 2016, pp. 391–402.

[3] S. Baltes, R. Kiefer, and S. Diehl, "Attribution required: Stack overflow code snippets in GitHub projects," in *39th International Conference on Software Engineering (ICSE 2017), Companion Volume*, S. Uchitel, A. Orso, and M. P. Robillard, Eds. Buenos Aires, Argentina: IEEE Computer Society, 2017, pp. 161–163.
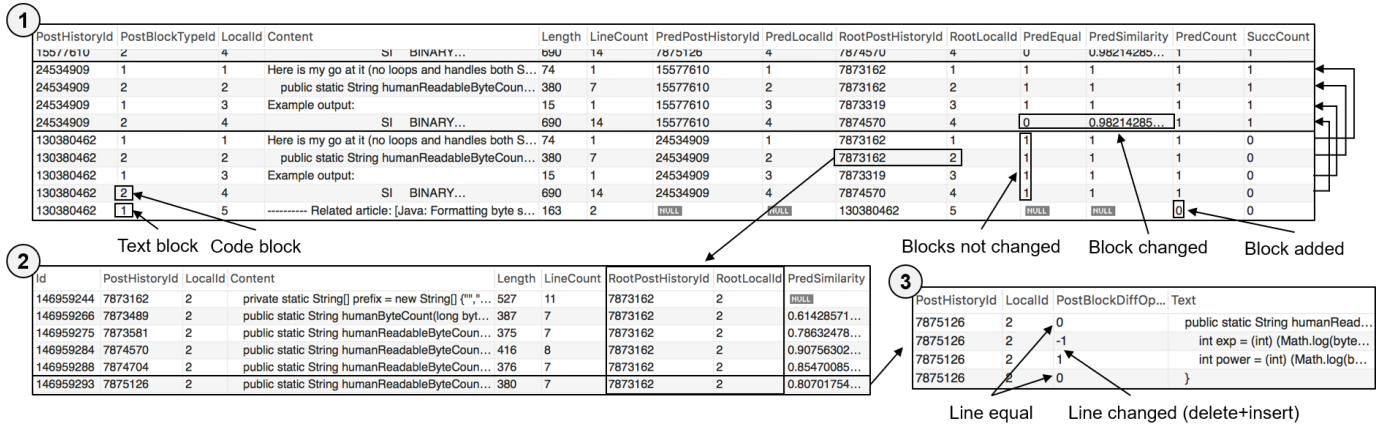
---

[2] https://stackoverflow.com/q/3758606

[3] http://empirical-software.engineering/blog/sotorrent

[4] http://empirical-software.engineering/blog/sotorrent-edithistory

Figure 3

PostHistoryId	PostBlockTypeId	LocalId	Content	Length	LineCount	PredPostHistoryId	PredLocalId	RootPostHistoryId	RootLocalId	PredEqual	PredSimilarity	PredCount	SuccCount
15577610	2	4	SI BINARY...	690	14	7875126	4	7874570	4	0	0.98214285...	1	1
24534909	1	1	Here is my go at it (no loops and handles both S...	74	1	15577610	1	7873162	1	1	1	1	1
24534909	2	2	public static String humanReadableByteCoun...	380	7	15577610	2	7873162	2	1	1	1	1
24534909	1	3	Example output:	15	1	15577610	3	7873319	3	1	1	1	1
24534909	2	4	SI BINARY...	690	14	15577610	4	7874570	4	0	0.98214285...	1	1
130380462	1	1	Here is my go at it (no loops and handles both S...	74	1	24534909	1	7873162	1	1	1	1	0
130380462	2	2	public static String humanReadableByteCoun...	380	7	24534909	2	7873162	2	1	1	1	0
130380462	1	3	Example output:	15	1	24534909	3	7873319	3	1	1	1	0
130380462	2	4	SI BINARY...	690	14	24534909	4	7874570	4	1	1	1	0
130380462	1	5	---------- Related article: [Java: Formatting byte s...	163	2	NULL	NULL	130380462	5	NULL	NULL	0	0

Text block   Code block   Blocks not changed   Block changed   Block added

Id	PostHistoryId	LocalId	Content	Length	LineCount	RootPostHistoryId	RootLocalId	PredSimilarity
146959244	7873162	2	private static String[] prefix = new String[] {"","...	527	11	7873162	2	NULL
146959266	7873489	2	public static String humanByteCount(long byt...	387	7	7873162	2	0.61428571...
146959275	7873581	2	public static String humanReadableByteCoun...	375	7	7873162	2	0.78632478...
146959284	7874570	2	public static String humanReadableByteCoun...	416	8	7873162	2	0.90756302...
146959288	7874704	2	public static String humanReadableByteCoun...	376	7	7873162	2	0.85470085...
146959293	7875126	2	public static String humanReadableByteCoun...	380	7	7873162	2	0.80701754...

PostHistoryId	LocalId	PostBlockDiffOp...	Text
7875126	2	0	public static String humanRead...
7875126	2	-1	int exp = (int) (Math.log(byte...
7875126	2	1	int power = (int) (Math.log(b...
7875126	2	0	}

Line equal   Line changed (delete+insert)

Fig. 3. Version data for Stack Overflow answer with ID 3758880 (truncated, based on tables `PostBlockVersion` and `PostBlockDiff`).

Figure 4

RepoName	Branch	Path	FileExt	Copies	PostId	SOUrl	GHUrl
FlakyTestDetection/jackr...	trunk	oak-commons/src/main/java/org/apache/jackrab...	.java	13	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/FlakyTestDet...
xdtianyu/ImageViewer	master	app/src/main/java/org/xdty/imageviewer2/utils/U...	.java	1	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/xdtianyu/Ima...
cleliameneghin/sling	trunk	contrib/extensions/tracer/src/main/java/org/apac...	.java	11	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/cleliameneghi...
Lucki/opsu	master	src/itdelatrisu/opsu/Utils.java	.java	8	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/Lucki/opsu/m...
opencb/opencga	develop	opencga-app/src/main/java/org/opencb/opencg...	.java	1	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/opencb/open...
weberjn/jFM	master	src/main/java/de/jwi/jfm/FileWrapper.java	.java	1	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/weberjn/jFM/...
headwirecom/sling	trunk	bundles/commons/log/src/main/java/org/apache...	.java	27	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/headwirecom...
roalva1/opencga	develop	opencga-app/src/main/java/org/opencb/opencg...	.java	1	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/roalva1/openca...
pglass/awspush-maven-...	master	src/main/java/Util.java	.java	1	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/pglass/awspu...
ffromm/sling	trunk	bundles/commons/log/src/main/java/org/apache...	.java	27	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/ffromm/sling/t...
apache/jackrabbit-oak	trunk	oak-commons/src/main/java/org/apache/jackrab...	.java	13	3758880	http://stackoverflow.com/a/3758880	https://raw.githubusercontent.com/apache/jackr...

Fig. 4. GitHub references to Stack Overflow answer 3758880 (table `PostReferenceGH`, truncated).

[4] L. An, O. Mlouki, F. Khomh, and G. Antoniol, "Stack Overflow: A Code Laundering Platform?" in *24th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER 2017)*, M. Pinzger, G. Bavota, and A. Marcus, Eds. Klagenfurt, Austria: IEEE Computer Society, 2017, pp. 283–293.

[5] D. Yang, P. Martins, V. Saini, and C. V. Lopes, "Stack Overflow in Github: Any Snippets There?" in *14th International Conference on Mining Software Repositories (MSR 2017)*, J. M. Gonzalez-Barahona, A. Hindle, and L. Tan, Eds. Buenos Aires, Argentina: IEEE Computer Society, 2017, pp. 280–290.

[6] M. Gharehyazie, B. Ray, and V. Filkov, "Some From Here, Some From There: Cross-Project Code Reuse in GitHub," in *14th International Conference on Mining Software Repositories (MSR 2017)*, J. M. Gonzalez-Barahona, A. Hindle, and L. Tan, Eds. Buenos Aires, Argentina: IEEE Computer Society, 2017, pp. 291–301.

[7] R. Abdalkareem, E. Shihab, and J. Rilling, "On code reuse from StackOverflow: An exploratory study on Android apps," *Information and Software Technology*, vol. 88, pp. 148–158, 2017.

[8] X. Xia, L. Bao, D. Lo, P. S. Kochhar, A. E. Hassan, and Z. Xing, "What do developers search for on the web?" *Empirical Software Engineering*, vol. 22, no. 6, pp. 3149–3185, 2017.

[9] F. Fischer, K. Böttinger, 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 (S&P 2017)*, K. R. B. Butler, Ú. Erlingsson, and B. Parno, Eds. San Jose, CA, USA: IEEE Computer Society, 2017, pp. 121–136.

[10] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky, "You Get Where You're Looking For: The Impact Of Information Sources on Code Security," in *2016 IEEE Symposium on Security and Privacy (S&P 2016)*, M. Locasto, V. Shmatikov, and Ú. Erlingsson, Eds. San Jose, CA, USA: IEEE Computer Society, 2016, pp. 289–305.

[11] M. M. Lehman, "Programs, life cycles, and laws of software evolution," *Proceedings of the IEEE*, vol. 68, no. 9, pp. 1060–1076, 1980.

[12] N. Chapin, J. E. Hale, K. M. Khan, J. F. Ramil, and W.-G. Tan, "Types of software evolution and software maintenance," *Journal of Software Maintenance*, vol. 13, no. 1, pp. 3–30, 2001.

[13] T. Mens and S. Demeyer, Eds., *Software Evolution*. Berlin, Germany: Springer, 2008.

[14] M. W. Godfrey and D. M. German, "The past, present, and future of software evolution," in *Frontiers of Software Maintenance (FoSM 2008)*, H. Muller, S. Tilley, and K. Wong, Eds. Beijing, China: IEEE, 2008, pp. 129–138.

[15] S. Baltes, L. Dumani, C. Treude, and S. Diehl, "SOTorrent: Reconstructing and Analyzing the Evolution Stack Overflow Posts," in *15th International Conference on Mining Software Repositories (MSR 2018)*, A. Zaidman, E. Hill, and Y. Kamei, Eds. Gothenburg, Sweden: ACM, 2018, pp. 319–330.

[16] Google Cloud Platform, "GitHub Data," 2018. [Online]. Available: https://cloud.google.com/bigquery/public-data/github

[17] G. Gousios, "The GHTorrent dataset and tool suite," in *10th International Working Conference on Mining Software Repositories (MSR 2013)*, T. Zimmermann, M. Di Penta, and S. Kim, Eds. San Francisco, CA, USA: IEEE, 2013, pp. 233–236.

[18] S. Baltes and L. Dumani, "SOTorrent Data Set Version 2018-08-28," 2018. [Online]. Available: http://doi.org/10.5281/zenodo.1406983

[19] ——, "SOTorrent BigQuery dataset 2018-08-28," 2018. [Online]. Available: https://bigquery.cloud.google.com/dataset/sotorrent-org: 2018_08_28

[20] ——, "sotorrent/metric-evaluation on GitHub," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1045823

[21] S. Baltes, "sotorrent/db-scripts on GitHub," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1116346

[22] ——, "SOTorrent: Reconstructing and Analyzing the Evolution of Stack Overflow Posts — Supplementary Material," 2018. [Online]. Available: http://doi.org/10.5281/zenodo.1201553

[23] ——, "Usage and Attribution of Stack Overflow Code Snippets in GitHub Projects — Supplementary Material," 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1148069

[24] Stack Exchange Community Wiki, "Database schema documentation for the public data dump and SEDE," 2018-02-27. [Online]. Available: https://meta.stackexchange.com/a/2678

[25] Stack Exchange Inc, "Stack Exchange Data Dump 2017-12-01," 2017. [Online]. Available: https://archive.org/details/stackexchange/