A StackExchange Dataset of Developer Questions Related to Checked-in Secrets in Software Artifacts

Setu Kumar Basak*, Lorenzo Neil†, Bradley Reaves‡ and Laurie Williams§
North Carolina State University, USA
Email: *sbasak4@ncsu.edu, †lcneil@ncsu.edu, ‡bgreaves@ncsu.edu, §lawilli3@ncsu.edu

***Abstract*—Throughout 2021, GitGuardian's monitoring of public GitHub repositories revealed a two-fold increase in the number of secrets (database credentials, API keys, and other credentials) exposed compared to 2020, accumulating more than six million secrets. To our knowledge, the challenges developers face to avoid checked-in secrets are not yet characterized. In our artifact, we provide a dataset containing 779 questions mined from three StackExchange sites asked by developers related to checked-in secrets from three StackExchange sites. In addition, we provide 434 accepted answers provided by the other users of StackExchange to mitigate the challenge of checked-in secrets.**

I. Artifact Description

In this section, we describe the purposes and details of our artifact.

A. Purpose of Research Artifact

As our artifact is comprised of questions asked by developers, future researchers and tool developers can further investigate to resolve the challenges developers face. For example, researchers and tool developers can find developers' challenges in sanitizing the version control history or publishing packages during deplopyment without secrets. In addition, they can also investigate the current solutions suggested by StackExchange users. However, our study identified that all the community's solutions are not secure. The chance of leaking secrets remains though developers apply the suggested solution. As a result, our artifact of questions and accepted answers will aid the researchers and tool developers in working on specific challenges to mitigate the secret sprawl [1], [2], [3], [4], [5].

B. Badge for Artifact

We claim **"Artifacts available"** badge for our artifact as we will make available our artifact for researchers and tool developers.

C. Data Description

Our artifact contains 694 questions from Stack Overflow [6], 40 questions from Information Security [7], and 45 questions from Software Engineering [8] sites. In total, our artifact contains 779 questions from the three sites of StackExchange spanning from September 2008 to December 2021. In addition, our artifact contains 390 answers from StackOverflow, 17 answers from Information Security, and 27 answers from Software Engineering sites. Each answer is an accepted answer provided by the StackExchange community to mitigate a specific challenge of checked-in secrets. Each field of our question artifacts is described in Table I.

TABLE I: An overview of our artifact

Field Name	Description
Id	An unique identifier of the question.
Title	The title of the question.
Body	The description of the question.
Tags	The tags related to the question such as "security", "git" and "key-management".
CreationDate	The date when the question is posted.
Score	The count of upvotes in the question.
ViewCount	The number of users who viewed the question.
AnswerCount	The total number of answers posted in the question.
CommentCount	The total number of comments posted in the question.
FavouriteCount	The total number of users who marked the question as favourite.
ClosedDate	The date when the community marked the question as closed.
URL	The url of the question.
AcceptedAnswerId	The unique identifier of the accepted answer for the question.
Answer	The accepted answer of the question.

D. Technology Skills

A reviewer who knows about StackExchange can verify our artifact.

E. Needed Softwares

To check the artifact, the reviewer will only need any software that can open the CSV and Excel files.

F. Artifact Storage

Our artifact is stored in Zenodo[1]. The artifact can also be accessed from our Github repository[2].

G. Ethics

The contents of all the Stack Exchange sites are under Creative Commons (CC BY-SA 3.0) license [9] with the following requirements: "You are free to: *Share* - copy and redistribute the material in any medium or format, *Adapt* - remix, transform, and build upon the material for any purpose, even commercially" [9].

H. License

The artifact will be available under MIT license [10].

[1]https://doi.org/10.5281/zenodo.7553203
[2]https://github.com/setu1421/ICSE-2023-Artifacts

I. Conclusion

Software relies heavily on the use of secrets for authentication and authorization, and the exposure of secrets is increasing each day. Our artifact will help the researchers and tool developers in expediting the research on software secret management.

ACKNOWLEDGMENT

This work was supported by National Science Foundation 2055554 grant.

REFERENCES

[1] M. Meli, M. R. McNiece, and B. Reaves, "How bad can it git? characterizing secret leakage in public github repositories." in *NDSS*, 2019.

[2] "Medical Data Leaked on GitHub Due to Developer Errors," https://threatpost.com/medical-data-leaked-on-github-due-to-developer-errors/158653, [Online; accessed Jan 15, 2023].

[3] "No need to hack when it's leaking," https://www.databreaches.net/wp-content/uploads/No-need-to-hack-when-its-leaking.pdf, [Online; accessed Jan 15, 2023].

[4] M. Jackson, "Uber Breach 2022 – Everything You Need to Know," https://blog.gitguardian.com/uber-breach-2022, [Online; accessed Jan 4, 2023].

[5] Security Magazine, "200,000 patient records exposed by hardcoded credentials and improper access controls," https://www.securitymagazine.com/articles/93182-000, 2020, [Online; accessed January 12, 2022].

[6] "Stack Overflow," https://stackoverflow.com, [Online; accessed January 3, 2022].

[7] "Information Security," https://security.stackexchange.com, [Online; accessed January 3, 2023].

[8] "Software Engineering," https://softwareengineering.stackexchange.com, [Online; accessed January 3, 2023].

[9] "Stack Overflow Creative Commons Data Dump," https://stackoverflow.blog/2009/06/04/stack-overflow-creative-commons-data-dump, [Online; accessed June 11, 2022].

[10] "The MIT License," https://opensource.org/licenses/MIT, [Online; accessed February 23, 2022].