Unveiling A Hidden Risk: Exposing Educational but Malicious Repositories in GitHub

Md Rayhanul Masud
UC Riverside
mmasu012@ucr.edu

Michalis Faloutsos
UC Riverside
michalis@cs.ucr.edu

ABSTRACT

Are malicious repositories hiding under the educational label in GitHub? Recent studies have identified collections of GitHub repositories hosting malware source code with notable collaboration among the developers. Thus, analyzing GitHub repositories deserves inevitable attention due to its open-source nature providing easy access to malicious software code and artifacts. Here we leverage the capabilities of ChatGPT in a qualitative study to annotate an educational GitHub repository based on maliciousness of its metadata contents. Our contribution is twofold. First, we demonstrate the employment of ChatGPT to understand and annotate the content published in software repositories. Second, we provide evidence of hidden risk in educational repositories contributing to the opportunities of potential threats and malicious intents. We carry out a systematic study on a collection of 35.2K GitHub repositories claimed to be created for educational purposes only. First, our study finds an increasing trend in the number of such repositories published every year. Second, 9294 of them are labeled by ChatGPT as malicious, and further categorization of the malicious ones detects 14 different malware families including DDoS, keylogger, ransomware and so on. Overall, this exploratory study flags a wake-up call for the community for better understanding and analysis of software platforms.

1 PROBLEM DEFINITION

Are GitHub repositories enabling the spread of malware? This is the question that motivates our work. GitHub is the most widely used open source software platform. There are more than 28M public repositories in GitHub [5], among them 7.5K repositories are identified to contain malware source code; according to a recent study [4]. It clearly indicates that public GitHub repositories can host malicious contents. As a result, malicious repos can be published in the following ways; (a) repositories are self-determined to be educational publishing proof-of-concept of vulnerabilities and exploits [2], and (b) sometimes they can intentionally contain malwares [3]. Another way is to do that, they may share malicious contents, but promoting as "for educational purpose only". We illustrate such an example of an educational GitHub repository in Figure 2(a) that shares source code of a ransomware application; however the educational intent does not prevent any malicious actor from using it in an unwanted manner. We refer to these repositories as **MalEdu** for the rest of the paper.

The problem we address here is the following: given a GitHub repository that is self-promoted as published for educational purpose only, how can we determine whether it is likely to be malicious? So, the input to the problem is GitHub, and the expected output is a set of MalEdu repos. The challenges include: (a) collecting educational repos, and (b) identifying the malicious ones among them.

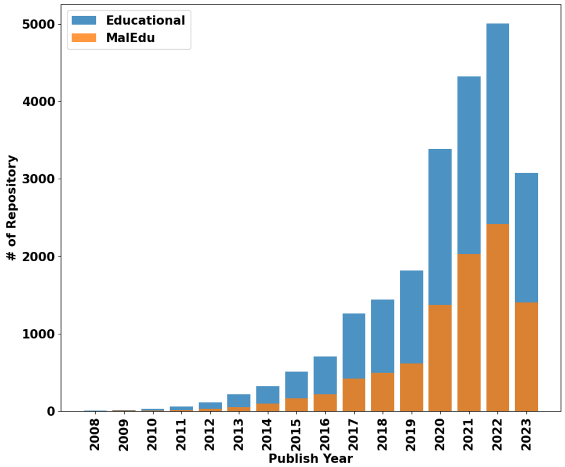


Figure 1: The number of educational GitHub repositories is increasing every year. The trend is similar for MalEdu (educational, but malicious) repositories.

2 CONTRIBUTION

As our key contribution, we propose a systematic study to analyze educational GitHub repositories to identify the repositories that contain malicious contents. We apply our method on a collection of 35.2K educational GitHub repositories (excluding forks) published during the period between 2008 and 2023 (Jun 24). Our key results are briefly discussed below.

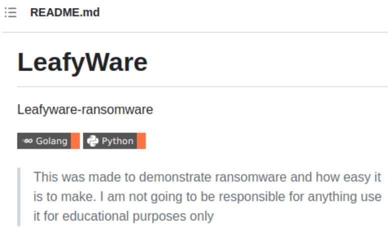
a. The number of educational repositories in GitHub has been increasing each year since its launch. According to figure 1, the frequency of the repositories published during 2020 and 2023 is 2.4 times the total number of repositories published before.

b. We find 35.2K educational repositories. 9294 (~26%) of them are identified as MalEdu repositories. Further categorization of MalEdu repos finds 14 different malware families.

c. Our manual validation suggests that ChatGPT accurately detects MalEdu repositories with 85% precision.

3 METHODOLOGY

A. Data collection. We use the GitHub search API to collect educational repositories. A GitHub repository has multiple metadata fields including repo title, description, readme file, star/fork/watch count and so on. We query the search API for repositories (excluding forks) with the following phrases in description and readme content; (a) education/educational purpose only, (b) only for education/educational purpose. This yields a collection of 35.2K educational repositories. Then, we filter the repos that contain both description and readme content, which results in 22.2K repositories that we consider in our experiment.



README.md

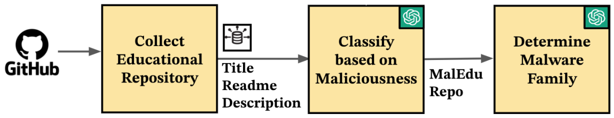
LeafyWare

Leafyware-ransomware

Golang Python

> This was made to demonstrate ransomware and how easy it is to make. I am not going to be responsible for anything use it for educational purposes only

(a) Example of a MalEdu GitHub Repository.



(b) Visualization of our approach.

Figure 2: (a) Example metadata of GitHub repository that hosts ransomware source code, while created for educational purpose only. (b) First, a collection of educational repos is classified by ChatGPT. Then, identified MalEdu repos are classified into malware families.

B. Determine maliciousness. Our aim is to determine how many educational repositories in GitHub platform can be labeled as malicious. A recent study [7] has found ChatGPT quite effective for answering health related questions (Yes/No) relying solely on the model knowledge. It motivates us to engage ChatGPT to classify the contents of a repository to identify the MalEdu repositories.

ChatGPT Prompt:
Context: Say you are a security professional. Given specific information about a repository, such as repo title, description and the readme file content, you will annotate the repository whether the repo is malicious.
User:
Repository Title: ...
Description: ...
Readme File Content: ...
Based on the provided information, please annotate with one option: benign, malicious, gray-area; indicating the potential maliciousness of the repository. No explanation needed.
ChatGPT: benign/malicious/gray-area.

We use ChatGPT API based on gpt-3.5-turbo-0613 to annotate the repositories in our dataset. We ask ChatGPT to choose one label among three categories; (a) benign, (b) malicious, and (c) gray-area; given title, readme content and description of a given repository indicating the maliciousness. Since ChatGPT is a generative model, we run two independent queries for each repository annotation, as suggested in a recent study [1]. Thus, we obtain two annotations from two queries for each of the repositories in our dataset. We extract the repositories which get a unanimous decision on the category “malicious" that provides us the list of MalEdu repositories.

C. Determine malware family. We want to determine the content type of the identified MalEdu repositories. To achieve this goal, first we create a list of popular malware families. Then, we ask ChatGPT to choose a family from the list for a given MalEdu repository. If the repository cannot be labeled using the list, ChatGPT is instructed to label it as “Miscellaneous". The detailed workflow is illustrated in Figure 2(b).

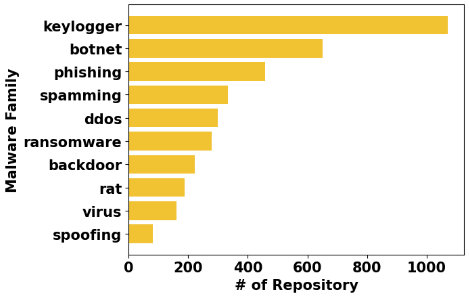


Figure 3: Top 10 malware families detected among MalEdu repos.

4 RESULTS AND EVALUATION

Results. We identify 9294 MalEdu repositories based on the annotations provided by ChatGPT. The normalized confusion matrix in figure 4 shows that ChatGPT annotations are found to be identical in almost all cases across both query processing.

We also detect 14 malware families during the categorization of MalEdu repository contents. We find “keylogger" as the most frequently identified malware family accounting to 1071 MalEdu repositories. Figure 3 lists top 10 malware families detected in this study.

Evaluation. To increase our confidence, we randomly select 100 “malicious" labeled (unanimously) GitHub repositories. Then, we investigate the contents of each of them for potential maliciousness. This manual investigation suggests that ChatGPT accurately detects MalEdu repositories with 85% precision.

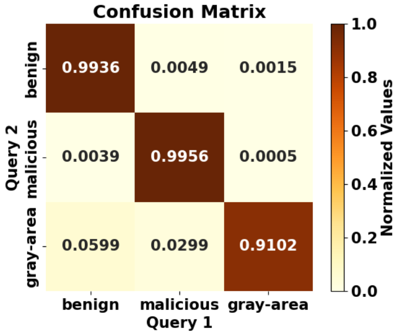


Figure 4: Normalized Confusion Matrix for ChatGPT annotations.

5 FUTURE WORK

Following the interesting findings, we intend to take a deep dive into the identified MalEdu repos for further profiling their authors and contents. We plan to investigate to detect any collaborative approach for the spread of such contents. In addition, we also want to verify the functionality of the source code to estimate the potential harm the MalEdu repositories can do.

6 RELATED WORK

Though several approaches aim to identify malware repositories in GitHub, none of them considers educational repositories as a possible source of malicious contents. A recent work, SourceFinder [4] gathers repositories based on keyword search, and then applies machine learning classifier on the repository content embedding to identify malware repos. Another recent study, GitCyber [6] incorporates cybersecurity domain knowledge along with code contents in a deep neural network for malicious repository detection.

REFERENCES

[1] Jialun Cao, Meiziniu Li, Ming Wen, and Shing-chi Cheung. 2023. A study on prompt design, advantages and limitations of chatgpt for deep learning program repair. *arXiv preprint arXiv:2304.08191* (2023).

[2] GitHub. 2023. https://docs.github.com/en/site-policy/acceptable-use-policies/github-active-malware-or-exploits

[3] The Hacker News. 2023. https://thehackernews.com/2023/06/fake-researcher-profiles-spread-malware.html

[4] Md Omar Faruk Rokon, Risul Islam, Ahmad Darki, Evangelos E Papalexakis, and Michalis Faloutsos. 2020. SourceFinder: Finding Malware Source-Code from Publicly Available Repositories in GitHub.. In *RAID*. 149–163.

[5] Wikipedia. 2023. https://en.wikipedia.org/wiki/GitHub

[6] Yiming Zhang, Yujie Fan, Shifu Hou, Yanfang Ye, Xusheng Xiao, Pan Li, Chuan Shi, Liang Zhao, and Shouhuai Xu. 2020. Cyber-guided deep neural network for malicious repository detection in GitHub. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE, 458–465.

[7] Guido Zuccon and Bevan Koopman. 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793* (2023).