Who is Creating Malware Repositories on GitHub and Why?

Nishat Ara Tania
University of California Riverside
Computer Science
ntani005@ucr.edu

Md Rayhanul Masud
University of California Riverside
Computer Science
mmasu012@ucr.edu

Md Omar Faruk Rokon
Walmart Global Tech
Advertisement Technology
mdomarfaruk.rokon@walmart.com

Qian Zhang
University of California Riverside
Computer Science
qzhang@cs.ucr.edu

Michalis Faloutsos
University of California Riverside
Computer Science
michalis@cs.ucr.edu

ABSTRACT

Recent studies have found thousands of malware source code repositories on GitHub. For the first time, we propose to understand the origins and motivations behind the creation of such malware repositories. For that, we collect and profile the authors of malware repositories using a three-fold systematic approach. First, we identify 14K users in GitHub who have authored at least one malware repository. Second, we leverage a pretrained large language model (LLM) to estimate the likelihood of malicious intent of these authors. This innovative approach led us to categorize 3339 as `Malicious`, 3354 as `Likely Malicious`, and 7574 as `Benign` authors. Further, to validate the accuracy and reliability of our classification, we conduct a manual review of 200 randomly selected authors. Third, our analysis provides insights into the authors' profiles and motivations. We find that `Malicious` authors often have sparse profiles and focus on creating and spreading malware, while `Benign` authors typically have complete profiles with a focus on cybersecurity research and education. `Likely Malicious` authors show varying levels of engagement and ambiguous intentions. We see our study as a key step towards understanding the ecosystem of malware authorship on GitHub.

CCS CONCEPTS

• **Security and privacy** → *Malware and its mitigation*; • **Computing methodologies** → *Supervised learning by classification*; • **Information systems** → *Web crawling*.

KEYWORDS

GitHub, Repository, Malware, User, Hacker, Classification, LLM

ACM Reference Format:
Nishat Ara Tania, Md Rayhanul Masud, Md Omar Faruk Rokon, Qian Zhang, and Michalis Faloutsos. 2024. Who is Creating Malware Repositories on GitHub and Why?. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651582

1 INTRODUCTION

GitHub is a widely known platform for collaborative software development [8], which has become the hub of open-source software development. However, such broad accessibility has made it also a fertile place for malicious activities. Researchers have identified the existence of thousands of malware repositories that are being shared openly [2, 13]. Here, the term **malware repository** refers to a repository that contains any software source code that can participate in compromising devices and supporting offensive, undesirable, and parasitic activities, such as a botnet or a keylogger. Given the prevalence of this phenomenon, it is natural to wonder who are the people that create these repositories.

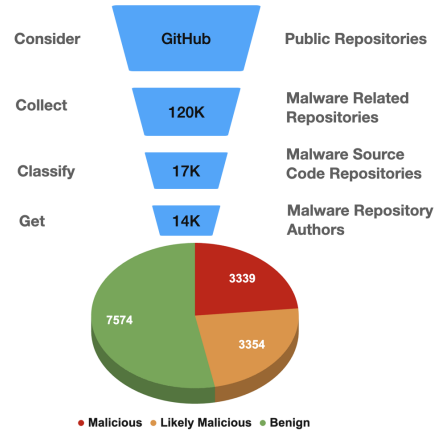

Figure 1: We have found 3339 malicious authors creating malware repositories on GitHub. Our large scale pipeline identified 120K malware repositories, which led to 17K malware repositories and 14K authors.

In response to this growing concern, we investigate the authors of malware repositories on GitHub. Specifically, we seek to answer—Who are the individuals that create malware repositories on GitHub, and what motivates them? To answer this question, we leverage their GitHub profiles, which contain a large number of mostly optional features. These features have a very wide range, which includes a bio, their twitter account, and whether they are looking for a job. In this work, the input to this problem consists of these user-profiles for authors of malware repositories, and the desired output is twofold: (a) whether the user is indeed malicious, and (b) a broader understanding of their intention and activities. Note that some security researchers create malware or malware-like repositories for educational purposes.

Prior works have made significant impacts in identifying and analyzing malware repositories [9, 13] and malicious GitHub users [4, 14, 15]. Early attempts to profile GitHub users [18] and understand community dynamics [11, 19] have laid the groundwork for our investigation. While these efforts have established the foundation, it is still necessary to get a deep understanding of the intentions of the authors of these malware repositories. This paper shifts the focus *from the repositories to the individuals* who create and maintain them. Additionally, the advent of Large Language Models has introduced new avenues for

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3651582

low-resource classification [10, 16], with ChatGPT demonstrating its effectiveness in text annotation and harmful content detection [3, 17].

Our contribution is a systematic study for classifying and profiling the authors of malware repositories.

A. Classification via Large Language Models (LLM): We employ OpenAI's ChatGPT-4 [12] for classifying authors by potential malicious intent, leveraging its proficiency in text classification and adaptive learning in few-shot or zero-shot contexts [6, 7]. Our methodology includes detailed prompt engineering to accurately classify authors into three distinct groups:

- `Malicious`: Authors whose activities are clearly harmful or intended for unauthorized access or damage.
- `Likely Malicious`: Authors whose profiles suggest a possibility of harmful intent, but without definitive evidence for a clear malicious classification.
- `Benign`: Authors engaged in benign activities, such as cybersecurity or educational research, without any harmful intentions.

B. Profiling and Validation: We provide a detailed analysis of 14K GitHub authors, revealing their characteristics and motivations. This involves manual investigation and validation of a subset of profiles to ensure the accuracy of our LLM-based classification. The key findings of applying our approach on the 14*K* authors are summarized below.

a. Accuracy of classification: We find that the LLM classification achieves 86% overall accuracy.

b. Distribution of authors: Our systematic approach identifies 3339 as `Malicious`, 3354 as `Likely Malicious`, and 7574 as `Benign` authors.

c. Profile characteristics and motivation: Our analysis highlighted that `Malicious` authors often have sparse profiles with low engagement, focusing on harmful activities. `Likely Malicious` authors have varied profiles and motivations, often straddling the line between curiosity and unethical activities. `Benign` authors exhibit complete profiles and actively contribute to cybersecurity research.

2 DATA COLLECTION METHODOLOGY

Our study builds on the foundation laid by SourceFinder [13] in 2020, which demonstrated high precision in identifying malware repositories on GitHub. Leveraging this methodology, we expanded our dataset to include not just repositories but also detailed profiles of the authors behind these repositories.

Identifying Malware Repositories and Authors We performed an extensive search for malware repositories, using 137 malware-specific keywords with the GitHub Search API in late 2022. Despite API restrictions limiting results to 1K entries/query, we overcame the limits by employing various criteria (e.g., most stars, fewest stars). This strategy enabled us to compile a comprehensive dataset of 120K potential malware-related repositories. Finally, we get 16726 malware repositories and 14267 malware authors using SourceFinder.

Profile Data Collection For each identified malware author, we collected their profile data using the GitHub Rest API, including username, full name, email, location, twitter username, short bio, blog url, company name, hireable status, account type (user/organization), list of followers, list of following, list of public repositories, creation date and last update date. Among them, full name, email, location, twitter username, short bio, blog url, company name, and hireable status are optional fields to be added by users.

Analyzing Author Engagement and Profile Completeness Our analysis reveals significant insights into the engagement levels and profile completeness of malware authors on GitHub. For instance, a majority of these authors tend to maintain a low profile with minimal social engagement, indicating a tendency towards privacy or anonymity. Furthermore, we observed correlations between certain profile attributes, suggesting patterns of information sharing among authors with more complete profiles. In Figure 2, we analyze the optional fields provided by authors in their GitHub profiles and the key findings are the following,

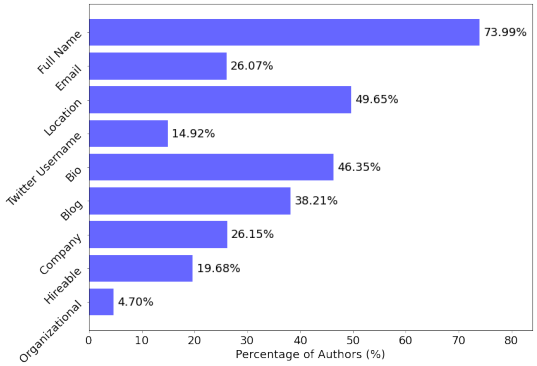

Figure 2: Percentage of malware authors including optional fields in their GitHub profiles, highlighting differences in profile completeness.

- The vast majority of malware authors operate a limited number of repositories, with a significant portion showing minimal social engagement on GitHub.
- Detailed analysis of optional profile fields reveals patterns in information sharing, with a disparity between authors with complete and sparse profiles.
- Authors with comprehensive profiles exhibit higher engagement levels, suggesting a correlation between profile completeness and community involvement.

This streamlined approach to data collection and analysis underscores the diversity and complexity of GitHub's ecosystem, particularly concerning the presence and activities of malware authors.

3 OUR CLASSIFICATION APPROACH

Our approach classifies GitHub profiles into three categories: `Malicious`, `Likely Malicious`, and `Benign`, based on an analysis of profile content. We employ OpenAI's ChatGPT-4 for this task, taking advantage of its demonstrated capabilities in text classification and its adaptability to diverse data scenarios. The model's efficiency in few-shot and zero-shot learning scenarios, as detailed in [7] and [6], suits our context of scarce labeled data. Furthermore, its prompt-based learning approach, as explored in [5], [10] and [1] suggest LLMs as viable alternatives to manual annotation.

To ensure the integrity and ethical consideration of our classification process, we meticulously followed these steps:

Step 1 - Ethical Consideration and Initial Context Establishment: Prior to classification, we ensure that our methodology is aligned with ethical guidelines, particularly concerning the potential implications of our findings with varied prompt engineering. We then provide ChatGPT-4 with a detailed context of our task, supplemented by annotated examples representative of each classification category to illustrate the nuances in user profile attributes.

Step 2 - Temperature-Varied Response Generation: The temperature parameter in LLMs like ChatGPT-4 influences the randomness of the generated text. A lower temperature yields more predictable and conservative responses, while a higher temperature produces a wider range of responses, including more creative or less common outputs. We systematically vary the temperature across five distinct

values: 0.0 (highly deterministic), 0.3, 0.5, 0.7, and 1.0 (highly creative), to capture a comprehensive response spectrum for each user profile.

Step 3 - Data Preparation: Each GitHub user profile is transformed into a structured prompt that includes relevant attributes such as username, bio, repository statistics, and other profile details. These prompts serve as input for ChatGPT-4 to evaluate and classify.

Step 4 - Classification Process: Using the aforementioned temperature settings, we submit each user profile prompt to ChatGPT-4 five times, each at a different temperature level. This approach is designed to gather a diverse set of perspectives on each profile, thereby reducing the likelihood of overfitting to a single pattern of response.

Few Shot ChatGPT Prompt:

role: system ***content:*** As an AI tasked with analyzing GitHub user profiles, your goal is to determine if users, known for creating malware repositories, are `Malicious`, `Likely Malicious`, or `Benign` where `Malicious` indicates clear harmful intent or unauthorized activities, unethical engagements, and `Likely Malicious` suggests potential harmful intent without conclusive evidence and `Benign` reflects legitimate, educational. Carefully review the following examples for guidance.
role: user ***content:*** Details for user1...
role: assistant ***content:*** Malicious
.
.
role: user ***content:*** Details for user6...
role: assistant ***content:*** Benign
role: user ***content:*** Evaluate the following user based on the above mentioned examples:
Details for user 'cacadosman': - Full Name: Fadli Maulana - Email: wetmanz23@gmail.com - Location: Indonesia - Blog URL: https://www.cacadosman.my.id/ - Bio: Nub InfoSec Engineer 100% human. - Company: GDP Labs - Twitter Handle: cacadosman - Hiring Status: Open to hiring opportunities - Profile Type: Individual User - Followers: 146 - Following: 123 - Public Repositories: 41 - Malware Repositories: 1 - Username Change: Never

Based on the provided information, please annotate with one option: `Malicious`, `Likely Malicious`, or `Benign`; indicating the potential maliciousness of the user. No explanation is needed.
ChatGPT: benign/malicious/gray-area.

Figure 3: ChatGPT-4 prompt for labeling maliciousness.

Step 5 - Majority Voting Mechanism: The multiple responses obtained for each profile at different temperatures are then subjected to a majority voting process. A list of five responses per user profile generated from five temperature settings is then subjected to a majority voting process. This entails selecting the most frequently occurring classification from the set of responses as the final verdict for each profile, thereby reinforcing the decision's robustness.

Step 6 - Prompt Engineering: Crafting effective prompts for ChatGPT-4 is a critical aspect of our methodology, involving several key steps to ensure accuracy and relevance in profile classification:

a. Selecting Profile Attributes: We identify user profile attributes, such as bios and repository counts, that provide insights into user behavior. These attributes inform their relevance and informativeness.

b. Balancing Prompt Detail: We aim for an optimal level of detail in our prompts, ensuring they are informative without overwhelming the AI. This balance is achieved through iterative refinement.

c. Refining Prompt Language: To minimize misinterpretation by ChatGPT-4, we carefully craft our prompts using clear and precise language. This step involves testing and adjusting the phrasing to improve the model's response accuracy.

d. Utilizing ChatGPT-4's Capabilities: By tailoring our prompts to align with the AI's operational strengths, we enhance the classification process's effectiveness. These steps ensure that our prompts effectively leverage ChatGPT-4's capabilities, facilitating accurate and insightful classification of GitHub user profiles.

Step 7 - Final Prompt Template and Ethical Approval: The final prompt template in Figure 3, designed for adaptability, underwent ethical review to ensure non-disclosure of sensitive information. This step underscores our commitment to ethical research practices and transparency in our methodology.

This comprehensive approach, from ethical considerations to final prompt optimization, ensures the integrity, reproducibility, and ethical compliance of our classification process.

4 EXPERIMENT AND RESULT

This section presents our analysis and classification of GitHub user profiles, applying the methodology from Section 3 to discern and profile authors of malware repositories. Our investigation sheds light on the diverse motivations and behaviors within the GitHub community, from malicious intent to constructive contributions in cybersecurity.

Final Classification Result: We classified 14267 GitHub authors, with 3339 identified as `Malicious` (indicative of harmful intent), 3354 as `Likely Malicious` (suggesting ambiguous intentions), and 7574 as `Benign` (focused on positive contributions to cybersecurity). These results underscore the complexity of user motivations on GitHub.

Validation: To ensure the accuracy of our classification, we conducted a manual review of a uniformly random sample of 200 authors. It includes 40 malicious, 45 likely malicious and 115 benign authors. This validation was carried out by three domain experts, each independently assessing the authors' profiles and GitHub activities. The experts were provided with guidelines to aid in their evaluation, ensuring consistency across classifications.

For each author, the domain experts assigned one of the three categories based on their judgment. The final label for each author was determined by a majority vote among the experts' classifications. This approach allowed us to mitigate individual bias and achieve a consensus on the most accurate categorization for each author.

Our findings from this manual validation process are as follows:

- **`Malicious` Accuracy:** 92.5%, with 37 out of 40 authors correctly identified, underscoring the effectiveness of our classification methodology.
- **`Likely Malicious` Accuracy:** 77.8%, with 35 out of 45 authors accurately classified, reflecting the inherent challenge in categorizing authors with ambiguous intentions.
- **`Benign` Accuracy:** 86.9%, with 100 out of 115 authors correctly identified, highlighting the model's ability to recognize constructive user engagements.

These validation results affirm the robustness of our approach, showing that GPT-4 can effectively differentiate among GitHub user profiles with a nuanced understanding of their intents and activities.

Profiling the Authors: Our investigation into the GitHub authors of malware repositories reveals distinct patterns in their behaviors and motivations, summarized in Table 1 and 2.

a. `Benign` *authors tend to provide all information.* `Benign` authors often have the most complete profiles, with significant differences observed across categories. Specifically, `Benign` authors have a higher tendency to list full names (95.74%) and locations (76.26%) compared to

Table 1: Distribution of Availability of User Profile Fields

Profile Field	Malicious(%)	Likely Malicious(%)	Benign(%)
Full Name	28.28	70.07	95.74
Email	3.51	10.03	43.00
Location	10.82	27.75	76.26
Twitter	2.16	4.03	25.15
Bio	15.80	23.55	69.61
Blog	8.49	14.74	61.54
Company	5.55	9.28	42.58
Hireable	3.42	11.76	30.23
Organizational	7.65	2.27	4.48

Table 2: Summary of GitHub Author Profiles

Type	Traits	Goals
Malicious	Limited profile data. Few followers/ following. Aggressive repo content.	Creating/spreading malware with minimal community interaction.
Likely Malicious	Varying profile detail. Mixed follower/ following counts. Ambiguous repo content.	Exploring cybersecurity with potential for unethical actions. Mixed community engagement.
Benign	Detailed profiles, active in community. Education-focused repos.	Enhancing cybersecurity through education and collaboration.

`Malicious` (28.28% and 10.82%, respectively) and `Likely Malicious` authors (70.07% and 27.75%, respectively).

b. `Benign` *authors have more public repositories.* `Benign` authors are more active, with 13.38% managing over 100 repositories, contrasting with `Malicious` (3.96%) and `Likely Malicious` authors' reserved presence aligning more closely with malicious authors.

c. Malicious authors do not have a wide range of followers. There is a marked difference in the follower base; `Benign` authors have broader follower bases, with 87.13% having at least 3 followers, while `Malicious` authors are more isolated, 58.59% having no followers.

d. Malicious authors are less likely to follow others. `Benign` authors are more socially engaged on GitHub, with a greater percentage following over 100 users (7.25%), compared to the more reserved `Malicious` (1.62%) and `Likely Malicious` authors.

e. Why do they create malware repositories on GitHub? Our investigation into the motivations for malware repository creation on GitHub considers bio texts, repo contents, and overall behaviors of authors.

- `Benign` Authors: Primarily motivated by constructive engagement in cybersecurity, focusing on educational and research activities. Key terms from their bios, such as 'Developer', 'Software Engineer', and 'Security Research', reflect this orientation. 16% authors explicitly mention 'educational purposes' in their repositories, underscoring their ethical approach.

- `Malicious` Authors: Driven to develop and distribute malware, showing minimal community engagement, with bios often containing aggressive terms like 'hack', 'ransom', 'ddos' and 'malware', highlighting their harmful focus. The mere 1.43% mentioning 'educational purposes' in their repositories aligns with their primary intent to create actual malware.

- `Likely Malicious` Authors: Exhibit a mix of curiosity-driven exploration and potential unethical leanings, with their level of engagement indicating a more ambiguous stance. Terms like 'bypass', 'experiment', and 'learning hacking' from their bios suggest a blend of educational interest and potential malintent. 4.68% authors mention keywords like 'educational purposes' in their repositories.

Note on Ethical Considerations and Dataset Sharing: In presenting our findings, we have consciously chosen not to disclose the names of individual GitHub authors. This decision is rooted in ethical considerations and security concerns. We have consciously chosen to obfuscate identifying information of individual GitHub authors. We encourage further research and discourse in cybersecurity with this dataset, available at DOI: 10.5281/zenodo.10806593, reflecting our dedication to responsible and secure information sharing.

5 CONCLUSION

We develop a systematic approach to classify and study authors that have created at least one malware repository on GitHub. Utilizing ChatGPT-4, we classify these GitHub authors into `Malicious`, `Likely Malicious`, and `Benign` categories and we offer new insights into their motivation in creating malware repositories. Our key findings indicate that `Malicious` authors exhibit harmful intentions with minimal community engagement, while `Benign` authors want to contribute positively to cybersecurity. `Likely Malicious` authors display a mix of technical curiosity and malicious intentions.

Our work sets the stage for future investigations into user behavior within open-source communities and highilights yet another potential application of LLMs.

6 ACKNOWLEDGMENT

This work was supported by the NSF SaTC grant No. 2132642.

REFERENCES

[1] Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint* (2023).
[2] Alejandro Calleja, Juan Tapiador, and Juan Cabalero. 2018. The malsource dataset: Quantifying complexity and code reuse in malware development. *IEEE Trans. on Info. Forensics and Security* 14, 12 (2018), 3175–3190.
[3] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056* (2023).
[4] Risul Islam, Md Omar Faruk Rokon, Ahmad Darki, and Michalis Faloutsos. 2021. Hackerscope: The dynamics of a massive hacker online ecosystem. *SNAM* (2021).
[5] Kazuaki Kashihara, Kuntal Kumar Pal, Chitta Baral, and Robert P Trevino. 2023. Prompt-Based Learning for Thread Structure Prediction in Cybersecurity Forums. *arXiv preprint arXiv:2303.05400* (2023).
[6] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329* (2022).
[7] Yinheng Li. 2023. Apractical SURVEY ON ZERO-SHOT PROMPT DESIGN FOR IN-CONTEXT LEARNING. (2023).
[8] Antonio Lima, Luca Rossi, and Mirco Musolesi. 2014. Coding together at scale: GitHub as a collaborative social network. In *ICWSM*.
[9] Md Rayhanul Masud and Michalis Faloutsos. 2024. Unveiling A Hidden Risk: Exposing Educational but Malicious Repositories in GitHub. arXiv:2403.04419 [cs.SE]
[10] Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv preprint* (2023).
[11] Behnaz Moradi-Jamei, Brandon L Kramer, J Bayoán Santiago Calderón, and Gizem Korkmaz. 2021. Community formation and detection on GitHub collaboration networks. In *IEEE/ACM ASONAM*. 244–251.
[12] OpenAI. 2023. GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf.
[13] 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 *23rd RAID*. USENIX, 149–163.
[14] Md Omar Faruk Rokon, Risul Islam, Md Rayhanul Masud, and Michalis Faloutsos. 2022. PIMan: A Comprehensive Approach for Establishing Plausible Influence among Software Repositories. In *2022 IEEE/ACM ASONAM*. IEEE.
[15] Md Omar Faruk Rokon, Pei Yan, Risul Islam, and Michalis Faloutsos. 2021. Repo2vec: A comprehensive embedding approach for determining repository similarity. In *ICSME*.
[16] Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? *arXiv preprint arXiv:2103.08493* (2021).
[17] Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint* (2023).
[18] Yu Wu, Jessica Kropczynski, Patrick C Shih, and John M Carroll. 2014. Exploring the ecosystem of software developers on GitHub and other platforms. In *CSCW*.
[19] Xiaoya Xia, Zhenjie Weng, Wei Wang, and Shengyu Zhao. 2022. Exploring activity and contributors on GitHub: Who, what, when, and where. In *APSEC*. IEEE.