CYBER SECURITY COOPERATIVE RESEARCH CENTRE

SCReeD Dataset Declaration Form

Dataset name:* Peekaboo

Dataset version:* 1

Dataset URL:* https://github.com/CSCRC-SCREED/Peekaboo

Creation date:* 17 April 2024

Last update:* 17 April 2024

Author(s):* Gaber, M., Ahmed, M., & Janicke, H.

Author(s) affiliation(s): Edith Cowan University

Author contact(s):* m.gaber@ecu.edu.au

Keywords:* Dynamic Binary Instrumentation, DBI, Malware Analysis, Feature Extraction, Evasive Malware, Sophisticated Malware

Description/background:* Cyber-attacks continue to evolve, increasing in frequency and sophistication where Artificial Intelligence (AI) is becoming essential in detecting modern malware. However, the accuracy of AI in malware detection is dependent on the quality of the features it is trained with. Static and dynamic analysis of malware is limited by the widespread use of obfuscation and anti-analysis techniques employed by malware authors, where if an analysis environment is detected the malware will hide its malicious behavior. However, Dynamic Binary Instrumentation (DBI) allows deep and precise control of the malware sample, thereby facilitating the extraction of authentic features from sophisticated and evasive malware. We developed Peekaboo, a DBI tool to defeat the anti-analysis techniques and extract authentic behavior from live malware samples. We collected 18,527 malware samples across ransomware, spyware, trojans, botnets, worms, Advanced Persistent Threats (APT) and post exploitation tools where every sample includes type, family, and variant information, for example Ransomware-WannaCry-SHA256. We also collected 1,973 benign software samples for analysis.This dataset contains the results for each sample, that were run for up to 15 minutes, to observe not only the anti-analysis techniques used but also its complete behavior. For each malware sample, the network traffic, every opcode that is executed and every evasive technique that is used are captured.

Dataset funding: N/A

Attribute details:* There are three main folders in the linked repository. The Peekaboo Data folder contains zip files of the timestamped raw json files extracted by Peekaboo for each sample and are organised by the malware family. There is also a csv file

generated with analysis.py for each family. The Peekaboo Network Traffic folder contains zip files of the .pcap files extracted by Peekaboo for every sample organised by family. The Python Scripts folder contains the Python scripts detailed below.

Intended target (if specified): N/A

Format:* JSON, CSV, PYTHON, PCAP

License:* Creative Commons Attribution 4.0 License/Open Access

Standard compliance: N/A

Type:* JSON

Size:* 650GB

Availability: Public

Data status: Available

Data provenance:*

Source computing infrastructure:* Live system

Accompanying program(s)/script(s): Yes

Software installer or VM for replication: N/A

Generated or captured via:* Hardware

Category/categories:* ML training data

Published in: Research Square (Preprint)

Open research question(s) (if any): N/A

Potential use case(s) or application area(s): Cyber Security

Data access control:* Global access

Data retention period:* N/A

Data validation/checksum:* Click or tap here to enter text.

GDPR compliance:* Yes

Consent: Yes

Ethics approval: N/A

Ethics considerations: N/A

* *denotes mandatory field*

☒ I confirm that I have read, understood, and agreed to the submission guidelines, policies, and submission declaration.

☒ I confirm that the contributors of the dataset have no conflict of interest to declare.

☒ I agree to take public responsibility for my dataset's contents.

Matthew Gaber

Signature of Corresponding Author Date: 04 June 2024

(signed on behalf of all contributors)