Crytojacking Classification based on Machine Learning Algorithm

Wan Nur Aaisyah Binti Wan Mansor
Faculty of Science and Technology
Universiti Sains Islam Malaysia
Nilai, Malaysia
aaisyahisya1923@gmail.com

Azuan Ahmad
Islamic Science Institute
Universiti Sains Islam Malaysia
Nilai, Malaysia
azuan@usim.edu.my

Wan Shafiuddin Zainudin
CyberSecurity Malaysia, Seri Kembangan, Selangor, Malaysia
wanshafi@cybersecurity.my

Madihah Mohd Saudi
Islamic Science Institute
Universiti Sains Islam Malaysia
Nilai, Malaysia
madihah@usim.edu.my

Mohd Nazri Kama
Fakulti Teknologi dan Informatik Razak
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
mdnazri@utm.my

ABSTRACT

The rise of cryptocurrency has resulted in a number of concerns. A new threat known as cryptojacking" has entered the picture where cryptojacking malware is the trend for future cyber criminals, who infect computers, install cryptocurrency miners, and use stolen information from victim databases to set up wallets for illicit funds transfers. Worst by 2020, researchers estimate there will be 30 billion of IoT devices in the world. Majority of the devices are highly vulnerable to simple attacks based on weak passwords and unpatched vulnerabilities and poorly monitored. Thus it is the best projection that IoT become a perfect target for cryptojacking malwares. There are lacks of study that provide in depth analysis on cryptojacking malware especially in the classification model. As IoT devices requires small processing capability, a lightweight model are required for the cryptojacking malware detection algorithm to maintain its accuracy without sacrificing the performance of other process. As a solution, we propose a new lightweight cryptojacking classifier model based on instruction simplification and machine learning technique that can detect the cryptojacking classification algorithm. This research aims to study the features of existing cryptojacking classification algorithm, to enhanced existing algorithm and to evaluate the enhanced algorithm for cryptojacking malware classification. The output of this research will be significant used in detecting cryptojacking malware attacks that benefits multiple industries including cyber security contractors, oil and gas, water, power and energy industries which align with the National Cyber Security Policy (NCSP) which address the risks to the Critical National Information Infrastructure (CNII).

CCS Concepts

• **Security and privacy→Intrusion/anomaly detection and malware mitigation→Malware and its mitigation**

Keywords

Cryptojacking; Classification; Machine learning; Malicious software; Cryptomining

1. INTRODUCTION

There are various types of malware such as virus that works to replicate itself relentlessly, it infects files and programs to destroy valuable data or cause irreparable damage, worms is a malicious code that work by copy itself and spread through the computer and the infected network will slow down, Trojan work by sneaks into victims computer and act as legitimate program, ransomware works by encrypting important data at the victim's host ask for ransom form the victim to recover their data. Cryptojacking on the other hand infect your computers, install cryptocurrency miners, and use stolen information from your databases to set up wallets for illicit funds transfers. The major problem of cryprojacking is it takes a high amount of computational power which is potentially most harmful among malicious malware.

Cryptojacking massively surged in 2017 but coin mining was been around for a while and BitcoinPlus.com was launched by Java Script code for pooled mining. In December 2017, cryptojacking activity was increase when more than 8 million cryptojacking was blocked by Symantec. Bitcoin also was worth approximately around US$30 in June 2011 and $6,000 in August 2018. In the end of 2017, Coinhive and Monero was marketed as alternative to generate revenue. Surprisingly the price of Monero coin was increase ten times valued before and hovering among $130.

Security analyst has evaluated cryptojacking as a significant emerging threat. Hackers use malwares to turn the stolen computing

ICCBN '20, April 15–18, 2020, Auckland, New Zealand
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7504-7/20/04…$15.00
https://doi.org/10.1145/3390525.3390537

power into digital coins. Newly discovered method of cryptojacking constantly appearing in the new daily. It is compatible everywhere whether on websites, servers, PCs or mobile [1].

Compared with other malware such as ransomware, cryptojacking guaranteed a profit and no active contact with victims. Cryptojacking needs very minimal interaction with victims and the task are done automatically, in secret. The main reason between the raise of cryptojacking attack is because it only require adding a snippet of JavaScript to a website or a malicious advertisement to utilize the victim's computational power. Users are very likely to be exposed with cryptojacking as they can be attacked through a browser, or by watching a video which. This cybercrime can lead to increases of monetary resource to the victim especially for the cost of electricity [2].

In previous work, there are a lot of malware detection approaches have been done to recognize malware [3]. Besides, another problem are lack depth analysis on cryptojacking malware especially in the classification model and lack of algorithm proposed for cryptojacking classification. To address the problem, it is great to come out with new algorithm that can detect cryptojacking activities. The aims of this paper are:

1. To study the features of existing cryptojacking classification algorithm
2. To classify cryptojacking based on machine learning algorithm.
3. To evaluate the cryptojacking malwareclassification.

2. RELATED WORKS

This section study and discuss various research related to cryptojacking detections and classifications. SEcure In-lined Script Monitors for Interrupting Cryptojacks (SEISMIC) is an automated method that modifies the incoming Wasm binary program so that it can be self-profile as they execute and detect echo of the cryptomining activity. The feasibility of semantic signature-matching was investigated by the author for robustly detecting the execution of browser-based cryptomining script that was implemented in Wasm. SEISMIC proposed detection by using semantic code features which is harder to obfuscate as they are fundamental to the miner computational purpose [4]. The result shows that mining and non-mining computations exhibit a huge different on behavioural pattern and this detection achieve 98% accuracy to detect cryptomining activities.

CMTracker is a behaviour-based detector which consists of two runtime profilers to automatically track Cryptocurrency Mining scripts and their related domain. The step that involve in CMTracker to detect cryptojacking website is firstly, the researcher introduces the dataset that use to conduct large-scale study. Next, illustrate two types behaviour based approach for dynamically discover cryptocurrency mining page. The website that was automatically identifies will go through further verification to determine whether indeed in cryptojacking website. The researcher concludes that if the webpage utilize more than 10% execution time for hashing, there is a good indicator of cryptojacking attack [2].

MalwarE Detection Using Statistical Analysis (MEDUSA) work focused on target system behaviour. System event logs, operations, networking and registry activities of applications will be monitored to generate the system normal profile and detect the suspicious activities. System-centric approach are proposed by the researcher by using deviation-based outlier detection model to quantify degree of deviation of system behaviour by continuously monitor various system artifacts on the target system. The advantage of MEDUSA is it does not need to be trained to whole family but it more focused on system artifact and feature to detect abnormalities [3].

CapsNet is Capsule Network which is a machine learning that was proposed by Hinton to imitate biological neural organization more closely. CapsNet adds structures called capsules to a convolutionary neural network and uses dynamic routing to connect capsules so that relative relationships between objects can be numerically represented as a pose matrix. Among other benefits, it can effectively recognize multiple objects even when they overlap. Experimental data shows the appealing performance of CapJack, with instant detection rates as high as 87% and 99% within an 11-second window [5].

BMDetector is Browser Mining Detector method which hooks Javascript in kernel source of Chrome Webkit. The author analyses browser heap snapshot data structure features and stack data after script execution using browser parse layer Hook key functions, extracts malicious miner dynamic behavior features. The works make use of Recurrent Neural Network (RNN-based) automatic detection which including pre-processing, sample generation and detection. RNN algorithm has good applicability for dynamic detection, detection and analysis after restoration of key code in the parsing layer and stream processing, and higher accuracy in the training set. The result for BMDetector is testing by experimental environment, experimental process analysis, functional test and performance test that show the final classifier accuracy of BMDetector prototype sytem is 93.04% [6].

3. METHODOLOGY

This chapter aims explain the workflow of methodology for cryptojacking classification.

3.1 Dataset

The dataset used for this study consist of 138,047 samples with the shape of the normal dataset is 41,323 samples and the shape of the malware dataset is 96,724 samples with all having 56 features as illustrated in Figure 1.

3.2 Feature Selection Phase

The features were chosen from numerous network features in the packet-level features. Above all, the main challenge in feature selection is finding the most relevant features that led to the highest true positive rate. A large number of features in the dataset should be filtered and refined. In addition, some features correlate to each other, and this hinders the cryptojacking malware classification process. Moreover, some features may contain redundant

information from other features. Redundant features increase computational time and reduce classification accuracy. The extracted features will be stored as a sequence of comma separated values (CSV) files.

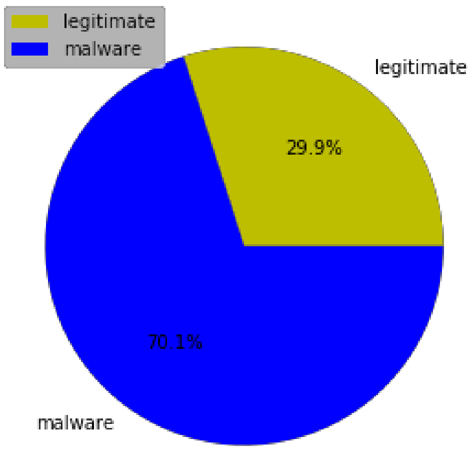

Figure 1. Distribution of dataset

3.3 Machine Learning Classifier Phase

In this stage, that is the machine learning classifier phase, the classifiers' output is produced. This phase determines the finest machine learning classifier for cryptojacking malware based on performance results.

3.4 Evaluation Phase

In this phase it will focus on evaluation of the algorithm for cryptojacking classification. In order to evaluate the performance, we used standard evaluation metrics which is False Positive Rate (FPR), False Negative Rate (FNR), True Positive Rate (TPR) and True Negative Rate (TNR).

4. RESULT AND DISCUSSION

This section will discuss the results that acquired from the experiment conducted using the collected datasets based on the methodology explained in previous section.

4.1 Feature Selection Phase

The dataset consists of two dimensional 138,047 entries of combination between malicious and benign application and 56 features had been extracted from each of the samples. From our study, 56 features are not feasible for machine learning training and testing thus feature selection should be perform to select best quality feature that fits with our machine learning training and testing.

In obtaining the best features, we used Extra Trees Classifier algorithm for this purpose. Extra Trees Classifier is an ensemble learning method fundamentally based on decision trees and applicable for feature selection. Based on the result, we able to reduce the features from 56 features to 13 best features based on the importance values produced by Extra Trees Classifier algorithm as showed in Table 1.

Table 1. Selected Features

Features	Importance Value
DllCharacteristics	0.12405669226638494
Characteristics	0.1229074401109358
SectionsMaxEntropy	0.09944400142308119
VersionInformationSize	0.09366836714692547
MajorSubsystemVersion	0.07301296036459513
Machine	0.06335850808839966
Subsystem	0.061749195616484766
ResourcesMinEntropy	0.04386165355695756
ImageBase	0.04234620214989661
ResourcesMaxEntropy	0.04005470252941394
SizeOfOptionalHeader	0.039088377638528224
MajorOperatingSystemVersion	0.03043914920802379
ResourcesMinSize	0.02149322187901585

4.2 Machine Learning Classifier Phase

This phase train and validate the dataset with selected features with two machine learning algorithms namely Random Forest and Gradient Boost algorithm. Table 2 summarize the confusion matrix for Random Forest and Table 3 summarizes confusion matrix for Gradient Boost Algorithm.

Table 2. Confusion Matrix for Random Forest Algorithm

	Predicted No	Predicted Yes
Actual No	19204	113
Actual Yes	76	8217

Table 3. Confusion Matrix for Gradient Boost Algorithm

	Predicted No	Predicted Yes
Actual No	19150	167
Actual Yes	183	8110

Table 3 shows the true positive and false positive rates for both algorithms.

Table 4. True Positive and False Positive Rates Result

	Random Forest	Gradient Boost
False Positive	0.5849769632965782	0.8645234767303411

5. ACKNOWLEDGMENT

This research financially support by the Universiti Sains Islam Malaysia Grant PPPI/UGC_0119/FST/051000/13519 and Universiti Teknologi Malaysia under UTMFR (Fundamental Research), vote no 21H12.

6. REFERENCES

[1] Carlin, D., OrKane, P., Sezer, S., & Burgess, J. (2018). Detecting Cryptomining Using Dynamic Analysis. In 2018 16th Annual Conference on Privacy, Security and Trust (PST) (pp. 1–6). IEEE. https://doi.org/10.1109/PST.2018.8514167

[2] Hong, G., Yang, Z., Yang, S., Zhang, L., Nan, Y., Zhang, Z., … Duan, H. (2018). How You Get Shot in the Back: A Systematical Study about Cryptojacking in the Real World. CCS '18 (ACM Conference on Computer and Communications Security), 13. https://doi.org/10.1145/3243734.3243840

[3] Ahmed, M. E., Nepal, S., & Kim, H. (2018). MEDUSA: Malware Detection Using Statistical Analysis of System's Behavior. In 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC) (pp. 272–278). IEEE.

[4] Wang, W., Ferrell, B., Xu, X., Hamlen, K. W., & Hao, S. (2018). SEISMIC: SEcure in-lined script monitors for interrupting cryptojacks. In Lecture Notes in Computer Science (including

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 11099 LNCS, pp. 122–142).

[5] Ning, R., Wang, C., Xin, C., Li, J., Zhu, L., & Wu, H. (2018). CapJack : Capture In-Browser Crypto-jacking by Deep Capsule Network through Behavioral Analysis, (April).

[6] Liu, J., Zhao, Z., Cui, X., Wang, Z., & Liu, Q. (2018). A novel approach for detecting browser-based silent miner. Proceedings - 2018 IEEE 3rd International Conference on Data Science in Cyberspace, DSC 2018, 490–497.