Improving Antivirus Signature For Detection Ransomware Attacks With Machine Learning

Alvian Bastian
e-mail: alvianbastian@poliupg.ac.id
Jurusan Teknik Elektro, Politeknik Negeri Ujung Pandang
Jl. Perintis Kemerdekaan Km. 10 Makassar, Indonesia

Abstrak

Kejahatan siber sulit dipisahkan dari perkembangan malware. Berdasarkan laporan dari Internet Security Threat, kejahatan dengan mengeksploitasi malware merupakan kejahatan yang cukup tinggi. Salah satu penyebaran malware yang cukup tinggi adalah ransomware. Infeksi akibat ransomware meningkat dari tahun ke tahun sejak 2013 dan terdapat 1.271 deteksi per hari selama tahun 2017. Sementara itu, pada tahun 2018 terjadi pergeseran serangan dimana 81 persen serangan menargetkan perusahaan sehingga infeksi ransomware menjadi 12 persen. Untuk mengatasi masalah ini, penelitian ini mengusulkan antivirus signature berdasarkan DLL Files dan API Calls untuk file ransomware. Mendeteksi file berdasarkan antivirus signature memiliki nilai teoritis dan praktik yang signifikan. Dari hasil percobaan menunjukkan pendeteksian file ransomware berdasarkan DLL Files dan functional API Calls dengan Machine Learning memiliki hasil yang baik dibandingkan mendeteksi file berdasarkan MD5 dan hexdump. Untuk pengujian dan deteksi ransomware files, penelitian ini menggunakan algoritma machine learning seperti KNN, SVM, Decision Tress, dan Random Forest. Hasil pengujian menunjukkan keberhasilan mendeteksi file ransomware, meningkatkan pendeteksian obyek, dan metode penelitian untuk antivirus signature.

Kata Kunci *: Ransomware, Antivirus, Machine Learning, Malware.*

1. Introduction

Internet has grown quickly. Data from International Telecommunication Union (ITU), there are 4.1 billion people are using internet in 2019. The global penetration rate increased from 16.8 % in 2005 to 53.6 % in 2019. Internet user grew on average by 10 % every year. In developed countries, 87 % people using the internet. (International Telecommunication Union, 2019) Rapid development and rapid internet growth and computer technology is followed by cybercrime activities. The Symantec Global Intelligence Network said there are 700.000 global adversaries and 98 million attack sensors worldwide. Symantec records 88.900 vulnerabilities from 24.560 vendors for 78.900 products. In 2015, email malware is increase from 1 in 220 emails, to 1 in 131 emails on 2016. Major emails are relying on first-stage downloaders like ransomware. The typical malware variants increase from 274 million in 2014 to 355 million in 2015, but largely stagnant in 2016 (0.5 percent increase from 2015 to 2016). Based on Symantec ISTR 2018, attacks from Ransomware have increased year by year since 2013 and peaked in 2016 at 1,271 detections for one day in 2017. But remained at those higher levels. With the amount of attacks 1,242 per day, it is nearly the similar with 2016. (Symantec, 2018) In 2018 there was a shift in attacks where 81 percent of attacks targeted enterprise so that ransomware infections increased by 12 percent. (Symantec, 2019) Antivirus signature is one way to prevent ransomware. (Wressnegger *et al.*, 2017) By doing detection of incoming file to computer can help during period of antivirus company release it update, so it becomes first aid when zero-day attacks. (Gardiner and Nagaraja, 2016) Portable Executables (PE) Ransomware files will be analyzed by static and dynamic analysis with using Pestudio and FileAlyzer, using open source antivirus ClamAV for build antivirus signature, and classification ransomware files with apply machine learning techniques for improve capability of antivirus signature. For extract DLL Files and functional API Calls, this experiment used module pefile in python and some ransomware is extracted using static analysis. (Kawaguchi and Omote, 2015)

The purposes of this research improve antivirus signature for detection Ransomware on host computer. The outcome of detection Ransomware will be improved with apply machine learning algorithms for get the top model deployment.

2. Methods

This section outlined the architecture of our system along with system overview and explain a workflow for detection Ransomware files.

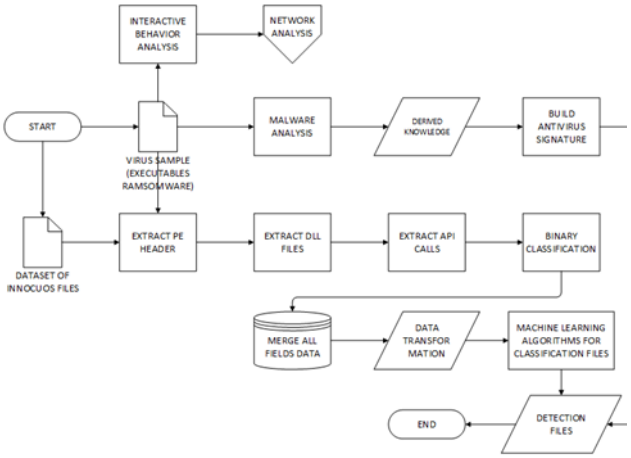

Fig. 1. Workflow diagram for Detection Ransomware Files

In workflow diagram to build antivirus signature and improve it with apply machine learning techniques for classification and detection Ransomware Attacks. Antivirus signature can detection ransomware files from its hash signature (MD5 and hex dump) and machine learning techniques are completing the system with classification and detection Ransomware files based on its DLL Files and API Calls. (Cabaj and Mazurczyk, 2016)

To analyze and facilitate in collecting the characteristics of files. This experiment uses pefile module on python. In the extraction section Portable Executables header (PE header) Ransomware aims to find Dos Header, File Header, and Optional Header so that the knowledge obtained from the characteristics of ransomware files. Some ransomware has anti-reverse engineering, so this research also used static analysis to extract its DLL Files and functional API Calls.

The PE format is a portable file format that can encapsulate the information required to manage executable code. Inside the file includes a reference library dynamic to connect import and export APIs. In windows, the PE format can be used for .EXE, .DLL, .SYS files, and so on. (Sebastián *et al.*, 2016)

With using Imported Address Table (IAT) as lookup table when application using application calls as function on different modules. Compiled programs do not know the memory location of the libraries, an indirect jump is required wherever API Call is made.

Dynamic linker will load and merge simultaneously, it will write the actual address in the IAT slot so that the memory location corresponds to library functions. Some viruses have the ability to hide from reverse engineering so call functions in Python cannot extract it. Therefore, we need Pestudio and FileAlyzer for extract its API Calls. (Al Amro and Alkhalifah, 2015)

After analyze the Ransomware files, this experiment should give label for each of record of Ransomware file to distinguish what is innocuous and malware file. Binary files will be used as controllers and runtime behaviors. In this section the extract process of some binary, construct feature, and classify samples to some parts of the malware and innocuous class. Some of the analyzed ransomware have similar behaviors to call some APIs with some arguments.

After merging all field data from ransomware files and innocuous files, data is transformed for presented dataset to machine learning algorithms. (Koret and Bachaalany, 2015)

A. Objectives and Dataset

The objectives of this experiments are built antivirus signature, classification and improve detection Ransomware files with using machine learning. The dataset of this experiments are 483 ransomware files from 8 class ransomwares i.e. Ransomware Badrabbit, Ransomware Cerber, Ransomware Locky, Ransomware Gandcrab, Ransomware Petya/NotPetya, Ransomware Sigma, Ransomware Tesla, and Ransomware WannaCry. The ransomware files are merged with 180 innocuous files to distinguish the characteristic of ransomware and innocuous files. The machine learning algorithms are Support Vector Machine, Decision Trees, Random Forest, and KNN. On the first experiment we analyze ransomware files to get MD5 and hex dump of each file. The MD5 and hex dump are base for build antivirus signature with using ClamAV, the open source antivirus. On the second experiment we analyze ransomware files to get DLL Files and API Call of each file. The DLL Files and API Call data that this experiment get from analyze ransomware files is used as input data for machine learning techniques. Table 1 is shown the sample files that is using on this research.

TABLE 1. THE SAMPLE FILES

ClassID	ClassName	Sum of Files
1.	Innocuous Files	180
2.	Badrabbit	68
3.	Cerber	72
4.	Locky	19

5.	GandCrab	53
6.	Petya/Not.Petya	62
7.	Sigma	70
8.	Tesla	67
9.	WannaCry	72
	Sum of Samples	663

B. Evaluation Method

For evaluation that present accuracy of detection ransomware files with using antivirus signature. To calculate accuracy with this method:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The following quantities are used:

TP = True Positive, FP = False Positive, FN = False Negative, FP = False Positive

To calculate the precision is given by:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity) is a percentage calculation of true positives, classify correctly about foreground regions, to calculate the recall is given by:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 Score is the weighted average of Precision and Recall. To calculate the F1 score is given by:

$$F1 - score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 * \frac{precision\ .recall}{precision + recall} \quad (4)$$

3. Results and Discussion

A. Antivirus Signature

The signature that is built on RemNux is sent to ClamAV for update its signature based on its MD5 and hexdump. The result is show ClamAV Antivirus success for detection Ransomware files.

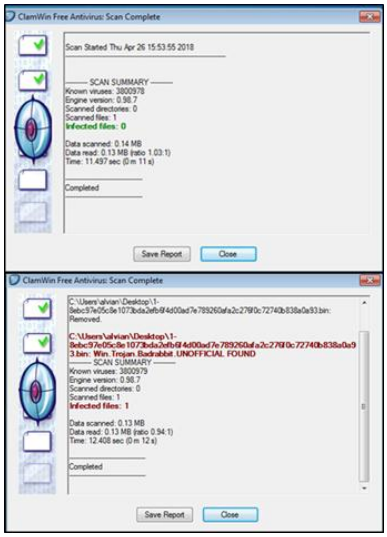

Fig. 2. Scanning Result before update (above) and after update (below) signature

B. Binary Classification

Data from Portable Executables (PE) Ransomware is extracted. The file includes a reference library (.DLL files) and API Calls. Selection DLL Files and API Calls based on the amount of frequencies that are called by ransomware files.

Each file is extracted and checked for each binary file for the binary file whether or not each feature of each string is present and then displayed in the vector. If the selected feature is available then it will be assigned a value of 1 otherwise it will be assigned a value of 0.

$$VectorSpace_{filex} = \begin{cases} 1 & if\ fi\ is\ in\ Filex \\ 0 & otherwise \end{cases} \quad (5)$$

Table 2 is shown the example binary vector space for detection Ransomware files. Table 2 is representative of dataset that is used for input Machine Learning Algorithm.

TABLE 2. EXAMPLE BINARY VECTOR SPACE AS DATASET

Class	KERNEL32.dll	USER32.dll	ADVAPI32.dll	…
innocuous	1	0	1	…
badrabbit	0	0	1	…
cerber	1	0	0	…
locky	0	0	1	…
gandcrab	1	1	1	…
petya	1	1	1	…
sigma	1	1	1	…
tesla	1	0	0	…
wannacry	1	0	0	…

C. Classification and Detection Ransomware Files

Data is trained using machine learning algorithms. This experiment using K-Nearest-Neighbors, Decision Trees, Random Forest, and Support-Vector-Machine.

Using cross-validation, choosing between models, and selecting features. The machine learning algorithms that is using on this experiment is supervised learning.

Using supervised learning aims to build models by generalizing out-of-sample data so that model evaluation procedures are possible to estimate how well the model is defined to perform on the out-of-sample data.

This experiment uses performance estimate to choose between available models. By doing each model train across the dataset and then evaluating each model by testing how well performs on the same data. This results in an evaluation metric known as training accuracy.

For testing and detection ransomware files, this research is using machine learning algorithms such as KNN, SVM, Decision Trees, and Random Forest. The result is shown on Table 3.

TABLE 3. CLASSIFICATION AND DETECTION RANSOMWARE FILES

	Precision	Recall	F1-score	Support
KNN	92%	91%	92%	199
Decision Trees	94%	93%	93%	205
Random Forest	92%	92%	92%	205
SVM	76%	75%	72%	199

The result of our experiment is shown, the Decision Trees has higher precision than other. The precision of Decision Trees is 94%, the precision of Random Forest and KNN is 92%. The result is shown that SVM is not good to implementation for detection Ransomware Files. For apply SVM to detection ransomware, we will need to reduce the sample data.

4. Conclusion

This article proposes Antivirus Signature for detection Ransomware Attacks. After experiments, and system testing, it shall be concluded as follows:

- Antivirus signature with MD5 and hex dump detection can discover the ransomware file which is same as MD5 and hex dump type. This is slightly effective for detecting malicious files with same MD5 and hex dump types but cannot detect malicious files with different MD5 and hex dumps. In fact, some ransomware files though the same variant but different MD5 and hex dump. This is why the build of antivirus signature based on MD5 and hex dump is not effective.
- To resolve inability of file detection based on MD5 and hex dump by antivirus signature. This experiment proposed machine learning techniques for classification and detection of ransomware files. For labeling and characterizing files based on DLL files and API calls that include on each file. The result is shown, our proposed schemed have 94 % for detection ransomware files.

5. References

[1] Al Amro, S. and Alkhalifah, A. (2015) 'A Comparative Study of Virus Detection Techniques', *International Journal of Computer, Electrical, Automation, Control and Information Engineering*.

[2] Cabaj, K. and Mazurczyk, W. (2016) 'Using software-defined networking for ransomware mitigation: The case of cryptowall', *IEEE Network*. doi: 10.1109/MNET.2016.1600110NM.

[3] Gardiner, J. and Nagaraja, S. (2016) 'On the security of machine learning in malware C&C detection: A survey', *ACM Computing Surveys*. doi: 10.1145/3003816.

[4] International Telecommunication Union (2019) 'Measuring digital development Facts and figures 2019', *ITUPublications*, pp. 1–15. Available at: https://www.itu.int/en/mediacentre/Documents/MediaRelations/ITU Facts and Figures 2019 - Embargoed 5 November 1200 CET.pdf.

[5] Kawaguchi, N. and Omote, K. (2015) 'Malware function classification using apis in initial behavior', in *Proceedings - 2015 10th Asia Joint Conference on Information Security, AsiaJCIS 2015*. doi: 10.1109/AsiaJCIS.2015.15.

[6] Koret, J. and Bachaalany, E. (2015) *The Antivirus Hacker's Handbook*, *The Antivirus Hacker's Handbook*. doi: 10.1002/9781119183525.

[7] Sebastián, M. *et al.* (2016) 'Avclass: A tool for massive malware labeling', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-45719-2_11.

[8] Rakhman, A., & Rais, R. (2020). Analisa Pakan Burung Otomatis Menggunakan Arduino Berbasis Internet Of Things. *Syntax Literate; Jurnal Ilmiah Indonesia*, *5*(5), 18-25.

[9] Symantec (2018) *Internet security threat report*, *Network Security*. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1353485805001947.

[10] Symantec (2019) 'Internet Security Threat Report VOLUME 21, February 2019', *Network Security*, 21(February), p. 61. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1353485805001947.

[11] Rakhman, A., & Sabanise, A. Y. F. (2019). Sistem Informasi Stok Kebutuhan Darah Pada Palang Merah Indonesia Dengan Metode Weighted Moving Average. *Syntax Literate; Jurnal Ilmiah*

Indonesia, *4*(7), 24-32.

[12] Wressnegger, C. *et al.* (2017) 'Automatically inferring malware signatures for anti-virus assisted attacks', in *ASIA CCS 2017 - Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*. doi: 10.1145/3052973.3053002.

P-ISSN :2089-676X
E-ISSN : 2549-0796