Combination method for malicious file detection

Muhammad Edzuan Zainodin[1,*], Rohayanti Hassan, Zalmiyah Zakaria[1], Shahreen Kasim [2]

[1]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, Batu Pahat, Johor, Malaysia

*muhammadedzuan@graduate.utm.my

ABSTRACT: With the advent of big data and cloud services, file security has become an important issue. Although a variety of detection and prevention technologies are used to protect file security, ransomware that demands money in exchange for one's data has emerged. This paper proposed xxx as a proof of concept to obtain the file entropy scoring and text analysis for file type identification to facilitate digital investigations in file type-based attacks.

Keywords: *Digital forensics; entropy; entropy scoring; file type identification;*

1. INTRODUCTION

Hackers and intruders continuously discover new vulnerabilities and attack computer users for various motives [1]. Users are generally aware of these threats and install firewall and antivirus software on computer system and keep their systems patched with latest software updates. Attackers bypass these defenses by sending malicious documents which exploit non-patched vulnerabilities in the application software. In this research work, we proposed of entropy calculation and n-gram analysis to the files and the proposed method can detect malicious documents. In past, Entropy was used to classify packed executables by Robert Lyda [2]. Degree of randomness in bytes of an encrypted executables should be high. In our research paper, belief is that degree of randomness in an exploit file should be less than genuine file of corresponding format. Researchers usually calculate entropy of their files under consideration. In [3], the researchers had done work with entropy of PDF files.

2. METHODOLOGY

Entropy is a method of measuring randomness or uncertainty in a given set of data. For calculating the entropy of file, our data set is sequence of bytes in the file [4]. Entropy can be calculated by using following formula

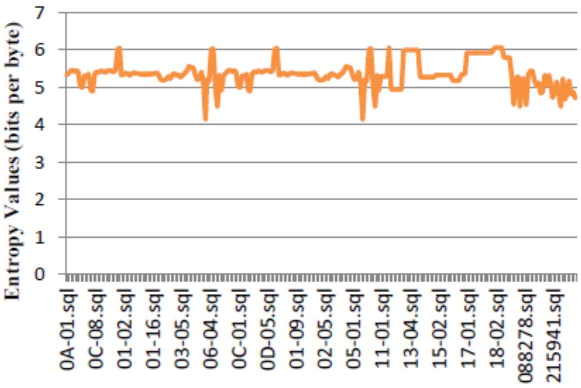
$$H(x) = -\sum_{i=1}^{N} p(i)\log_2 p(i)$$

Where *p(i)* is the probability of ith unit of information and specifically in our case it is ith byte of file. The value of H(x) will be high if probabilities of occurrence of bytes are low and vice versa. Encrypted Files can be classified using such entropy analysis with understanding that probabilities of occurrence of bytes will be low.

Meanwhile, an n-gram is a subsequence of N consecutive tokens in a stream of tokens [5]. This content-based approach looks for the binary content of the set of files and produce normalized n-gram distributions representing all files of a specific type. N-gram model able to determine the validity of files based on certain file types. It is also able to determine the type of an unnamed object. Thus, it is useful for file type identification due to its ability to present the order of bytes. However, n-gram models are often criticized because they lack any explicit representation of long range dependency. Due to this reason, n-gram models have not made much impact on linguistic theory, where part of the specific goal is to model such dependencies.

3. RESULTS AND DISCUSSION

This section presents results for some types of files. Three line graphs are plotted to visualize the change in entropy values for the three selected file types. Line graph is useful in laboratory research involving a correlation between different file entropy values. A conclusion can be drawn easily with a single observation based on the analysis of line graph.



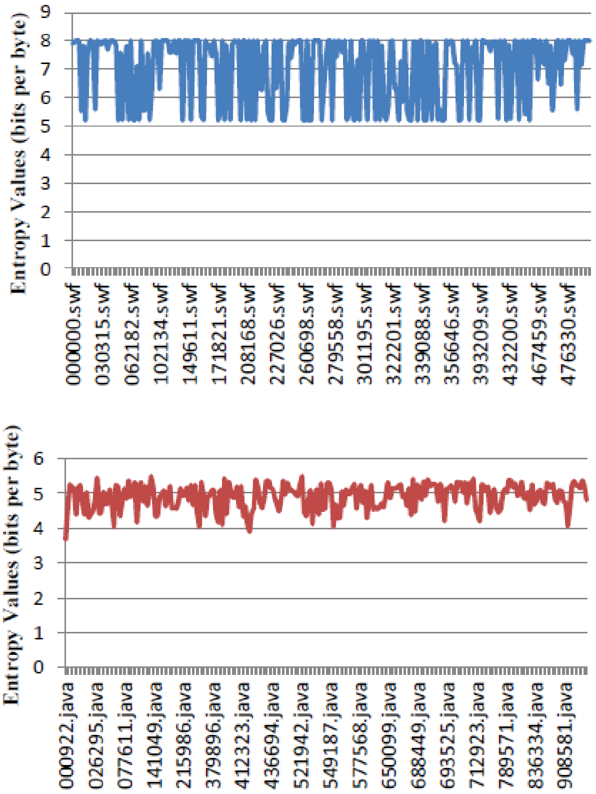


Figure 1 Preliminary results for entropy file detection

4. CONCLUSION

In general, entropy scoring is useful as a kind of statistical information to help forensic investigators reduce the amount of data required to be analyzed. For example, when the file entropy value is detected as low entropy, the contents of the file is considered as low randomness as the number of overwritten data is high and therefore forensic investigators can selectively pinpoint to the file types that contain high probability of digital evidences. They can obtain valuable information in a short period of time and use the statistics acquired to make discoveries, decisions and predictions based on data. The extended measurements of entropy for many files type allow forensic investigators to quickly limit their focus on particular units of information based on the specific target.

ACKNOWLEDGEMENT

This study was supported by Universiti Teknologi Malaysia under UTMER FASA 1/2020 (VOT NO 19J59).

REFERENCES

[1] Quick Heal Technologies Report, January 2013, available at http://www.quickheal.com/blog/news20/

[2] R. Lyda and J. Hamrock, “Using Entropy Analysis to Find Encrypted and Packed Malware,” IEEE Security and Privacy, March/April 2007

[3] Brandon Dixon, blog.9bplus.com

[4] Mike Schiffman, “Cisco Blog- Information Entropy”,http://blogs.cisco.com/security/on_inform ation_entropy.

[5] Li, W. J., Wang, K., Stolfo, S. J., & Herzog, B. (2005, June). Fileprints: Identifying file types by n-gram analysis. In Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop (pp. 64-71). IEEE.