SFEM: Structural Feature Extraction Methodology for the Detection of Malicious Office Documents Using Machine Learning Methods

Aviad Cohen, Nir Nissim, Lior Rokach, Yuval Elovici
Email: *{aviadd,nirni,liorrk,elovici}@post.bgu.ac.il*
Department of Information Systems Engineering, Ben-Gurion University of the Negev, Israel
Malware Lab, Cyber Security Research Center, Ben-Gurion University

ABSTRACT

Cyber-attacks aimed at organizations have increased since 2009, with 91% of all organizations hit by cyber-attacks in 2013.[1] Attacks aimed at organizations usually include harmful activities such as stealing confidential information, spying and monitoring an organization, and disrupting an organization's actions. Attackers may be motivated by ideology, criminal intent, a desire for publicity, and more. The vast majority of organizations rely heavily on email for internal and external communication. Thus, email has become a very attractive platform from which to initiate cyber-attacks against organizations. Attackers often use social engineering[2] in order to encourage recipients to click on a link which refers to a malicious website or opens a malicious attachment. According to Trend Micro,[3] attacks, particularly those against government agencies and large corporations, are largely dependent upon spear-phishing[4] emails.

Non-executable files such as Office or PDF documents attached to an email are a component of many recent cyber-attacks. This type of attack has grown in popularity, because of the filtering process of email servers; executable files (e.g. *.exe) attached to emails are filtered out by most email servers due to the risk they pose, while non-executables attachments are not filtered and are considered safe by most users. Non-executable files are written in a format that can be read only by a program that is specifically designed for that purpose and often cannot be directly executed. Unfortunately, non-executable files are as dangerous as executable files, since their readers can contain vulnerabilities that, when exploited, may allow an attacker to perform malicious actions on the victim's computer. Cybercriminals launch attacks through Microsoft Office files,[5] taking advantage of the fact that Office documents are widely used among most organizations; in fact, Microsoft Office's market share has held steady at 94% for years, with 500 million customers.[6] Cybercriminals exploit the fact that most employees within organizations do not take precautions when receiving and opening these files. The Symantec Internet Security Threat Report[7] reveals that Microsoft Office document file attachments have surpassed executable files as the most frequently used type of attachments in spear-phishing attacks.

To prevent such cyber-attacks, defensive tools such as firewalls, intrusion detection systems (IDSs), intrusion prevention systems (IPSs), anti-viruses, and others are used; however, these tools are limited in the detection of attacks that are launched via non-executable files, particularly when a sophisticated advanced persistent threat (APT) attack is executed against an organization. The main limitation of most existing detection tools lies in their inability to detect new unknown types of attacks based on known attack signatures, due to the time lag that exists between when a new unknown malware appears and the time anti-virus vendors update their clients with the new signature. During this period of time, many computers are

[1] http://www.humanipo.com/news/37983/91-of-organisations-hit-by-cyber attacks-in-2013/

[2] http://searchsecurity.techtarget.com/definition/social-engineering

[3] http://www.infosecurity-magazine.com/view/29562/91-of-apt-attacks-start-with-a-spearphishing-email/

[4] http://searchsecurity.techtarget.com/definition/spear-phishing

[5] http://securelist.com/blog/research/65414/obfuscated-malicious-office-documents-adopted-by-cybercriminals-around-the-world/

[6] http://www.dailytech.com/Office+2010+to+Launch+Today+Microsoft+Owns+94+Percent+of+the+Market/article18360.htm

[7] https://www4.symantec.com/mktginfo/whitepaper/ISTR/21347932_GA-internet-security-threat-report-volume-20-2015-social_v2.pdf

vulnerable to the new malware [1][1], [2]. The risk grows when the malware exploits an unknown vulnerability (zero-day).

Duqu, discovered on September 1, 2011 by CrySyS Lab,[8] is an infamous sophisticated cyberespionage malware thought to be related to the famous Stuxnet[9] APT worm. The Duqu malware looked for information that could be useful in attacking industrial control systems (e.g., SCADA). Duqu exploited a couple of zero-day vulnerabilities in order to operate, one of which was located in the Microsoft Word TrueType font parsing engine which allows the execution of arbitrary code.

Ransomware is a part of a recent malware trend aimed at individuals and organizations that prevents or limits access to resources in the infected computer [3], [4]. The Ransomware demands a ransom (payed to the malware operators) in order to remove the restriction. CryptoWall is a well-known ransomware which encrypts the host's files using a strong encryption algorithm, thus preventing access to the files. As of the end of 2015, CryptoWall has extorted approximately $352,000,000 from tens of thousands of victims worldwide. The victims include both businesses and individuals, many of whom are based in North America. [10] Ransomware is typically spread through emails which contain an attachment that, when opened, infects the computer. Ransomware has recently been observed in Office documents as well. [11, 12]

In this study, we present a novel structural feature extraction methodology (*SFEM*) that extracts discriminative structural features from Extensible Markup Language (XML) based documents (e.g., *.docx, *.xlsx, *.pptx, *.odt, *.ods, etc.). To the best of our knowledge, we are the first to present a feature extraction methodology tailored to XML-based documents. The extracted features contribute to the discrimination between malicious and benign documents when used in conjunction with machine learning algorithms. *SFEM* is aimed at enhancing the detection of malicious, XML-based documents. We demonstrate and evaluate the power of *SFEM* on the detection of Microsoft Word files (*.docx) and compare its performance against existing leading anti-virus engines.

We evaluated *SFEM* using a large and representative collection of Microsoft Word XML-based documents (*.docx) which contains 830 malicious and 16,180 benign files, and through three comprehensive experiments.

The first experiment designed find which configuration of feature selection, feature representation, top-feature selection, and classifier provides the best detection results. We considered the *Information Gain* and *Fisher Score* feature selection methods; *Boolean* and *TFIDF* feature representation methods; the following top-feature selection: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000 and 2,000; and the following machine learning classifiers: *Naïve Bayes*, *Bayes Network*, *J48*, *Random Forest*, *Logistic Regression*, *LogitBoost*, *SMO*, *Bagging,* or *AdaBoost.* The configuration that provides the best detection measures is based on: *TFIDF, Fisher Score, Top 200,* and *Random Forest (500 trees)*, and achieved a TPR of 0.97 with an FPR of 0.049, and an AUC of 0.9912. Figure 1 presents the ROC curve of the *Random Forest (500 trees)* in the best configuration. The X-axis represents the FPR, and the Y-axis represents the TPR. The area pictured in the red rectangle is enlarged and presented within.

[8] https://www.crysys.hu/

[9] http://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet

[10] http://www.darkreading.com/endpoint/with-$325-million-in-extorted-payments-cryptowall-3-highlights-ransomware-threat/d/d-id/1322899

[11] http://arstechnica.com/security/2016/02/locky-crypto-ransomware-rides-in-on-malicious-word-document-macro/

[12] http://thehackernews.com/2016/02/locky-ransomware-decrypt.html

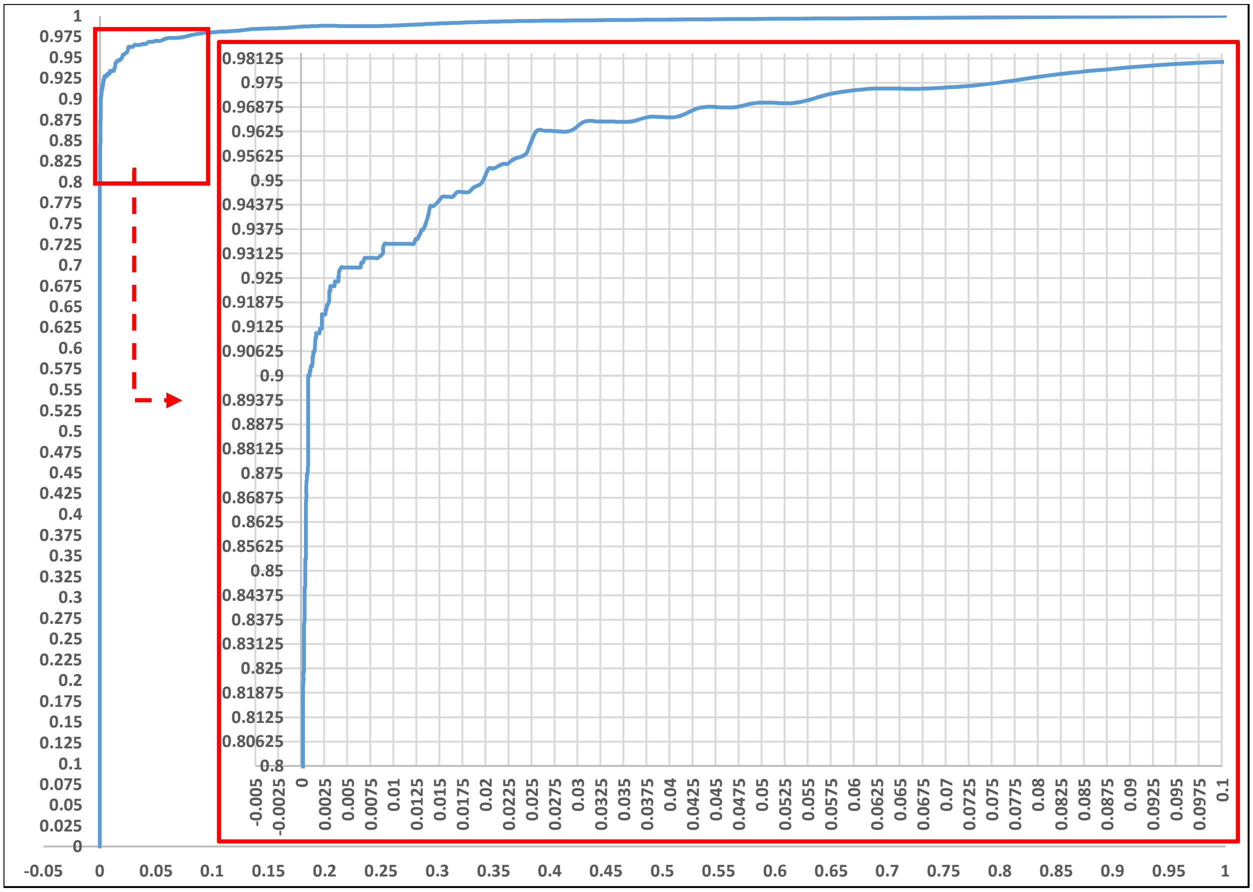

Figure 1. ROC curve of the Random Forest (with 500 trees) classifier applied on a dataset containing the top 200 features extracted by SFEM, selected by Fisher Score, and in TFIDF representation.

The second experiment design to determine whether removing the numbers from the features enhance the detection results (*SFEM No Number*). We found that removing the number from the extracted features significantly reduces the time complexity and computational resources needed for the feature extraction and selection processes, however it does not lead to significantly better detection results. When considering both *SFEM* and *SFEM NN*, the best configuration is based on: *SFEM NN, Fisher Score, TFIDF, top 900,* and *Random Forest (500 trees).* The classifier achieves a TPR of 0.97 with an FPR of 0.049, and an AUC of 0.9927.

The third experiment designed to compare our best configuration to top leading, well-known, anti-virus engines that are used by organizations. We found that the best configuration discovered in the first and second experiments significantly outperforms the top, leading anti-virus engines in the task of malicious *.docx detection. The AVAST anti-virus engine achieved a TPR of only 0.777, while our best configuration achieved a TPR of 0.97 (~25% better).

Given the challenges faced by organizations and cloud services it is clear that a more comprehensive detection method for malicious documents is needed. *SFEM* is static, light, and fast, and in conjunction with machine learning classifiers it offers an advanced detection model for known and unknown malicious XML-based office documents. Thus, it would be valuable to integrate that detection model in organizations and cloud services (e.g., Microsoft Office 365, Google Drive, etc.) in order to safeguard such networks and storage systems from malicious documents. Moreover, since such a detector is light and fast it can also be integrated into a Microsoft Office product.

REFERENCES

[1] M. Christodorescu and S. Jha. Testing malware detectors. ACM SIGSOFT Software Engineering Notes 29(4), pp. 34-44. 2004.

[2] S. R. White, M. Swimmer, E. J. Pring, W. C. Arnold, D. M. Chess and J. F. Morar. Anatomy of a commercial-grade immune system. IBM Research White Paper 1999.

[3] Kharraz, A., Robertson, W., Balzarotti, D., Bilge, L., & Kirda, E. (2015). Cutting the gordian knot: a look under the hood of ransomware attacks. In*Detection of Intrusions and Malware, and Vulnerability Assessment* (pp. 3-24). Springer International Publishing.

[4] Pathak, P. B., & Nanded, Y. M. A Dangerous Trend of Cybercrime: Ransomware Growing Challenge.