Expert Systems With Applications 63 (2016) 324–343

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

SFEM: Structural feature extraction methodology for the detection of malicious office documents using machine learning methods

Aviad Cohen [a,b,*], Nir Nissim [a,b], Lior Rokach [a,b], Yuval Elovici [a,b]

[a] Department of Information Systems Engineering, Ben-Gurion University of the Negev, Israel
[b] Malware Lab, Cyber Security Research Center, Ben-Gurion University, Israel

ARTICLE INFO

Article history:
Received 25 April 2016
Revised 16 June 2016
Accepted 5 July 2016
Available online 9 July 2016

Keywords:
Machine learning
Malware detection
Static analysis
Structural features
Microsoft office open xml
Document

ABSTRACT

Office documents are used extensively by individuals and organizations. Most users consider these documents safe for use. Unfortunately, Office documents can contain malicious components and perform harmful operations. Attackers increasingly take advantage of naive users and leverage Office documents in order to launch sophisticated advanced persistent threat (APT) and ransomware attacks. Recently, targeted cyber-attacks against organizations have been initiated with emails containing malicious attachments. Since most email servers do not allow the attachment of executable files to emails, attackers prefer to use of non-executable files (e.g., documents) for malicious purposes. Existing anti-virus engines primarily use signature-based detection methods, and therefore fail to detect new unknown malicious code which has been embedded in an Office document. Machine learning methods have been shown to be effective at detecting known and unknown malware in various domains, however, to the best of our knowledge, machine learning methods have not been used for the detection of malicious XML-based Office documents (*.docx, *.xlsx, *.pptx, *.odt, *.ods, etc.). In this paper we present a novel structural feature extraction methodology (*SFEM*) for XML-based Office documents. *SFEM* extracts discriminative features from documents, based on their structure. We leveraged *SFEM*'s features with machine learning algorithms for effective detection of malicious *.docx documents. We extensively evaluated *SFEM* with machine learning classifiers using a representative collection (16,938 *.docx documents collected "from the wild") which contains ~4.9% malicious and ~95.1% benign documents. We examined 1,600 unique configurations based on different combinations of feature extraction, feature selection, feature representation, top-feature selection methods, and machine learning classifiers. The results show that machine learning algorithms trained on features provided by *SFEM* successfully detect new unknown malicious *.docx documents. The Random Forest classifier achieves the highest detection rates, with an AUC of 99.12% and true positive rate (TPR) of 97% that is accompanied by a false positive rate (FPR) of 4.9%. In comparison, the best anti-virus engine achieves a TPR which is ~25% lower.

1. Introduction

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 the user 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.

* Corresponding author.
E-mail addresses: aviadd@post.bgu.ac.il, aviadjo@gmail.com (A. Cohen), nirni@post.bgu.ac.il (N. Nissim), liorrk@post.bgu.ac.il (L. Rokach), elovici@post.bgu.ac.il (Y. Elovici).

[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

http://dx.doi.org/10.1016/j.eswa.2016.07.010
0957-4174/

The well-known breach of the RSA Security[5] (acquired by EMC[6]) network in 2011 was carried out by attackers who sent phishing emails to two small groups of employees. A malicious Microsoft Excel file which contained malware was attached to the email. When the employees opened the Excel file, the malware exploited a vulnerability which allowed the attackers to gain control of machines and access servers in RSA's network. Recently, Kaspersky Lab[7] discovered a massive cyber-attack on the banking industry.[8] Many banks around the world were targeted by this sophisticated attack which aimed at penetrating the banks' secured systems and stealing their money. The initial breach in this case can be traced back to a Microsoft Word file containing malicious embedded code, again attached to an email. The breach affected more than 100 banks in 30 different countries worldwide and could total as much as $1 billion in total losses.

Non-executable files such as Office or PDF documents attached to an email are a component of many recent cyber-attacks, including those described in the previous paragraph. 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 only be read 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,[9] 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.[10] 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[11] 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 vulnerable to the new malware (Christodorescu & Jha, 2004), (White, Swimmer, Pring, & Arnold, 1999). The risk grows when the malware exploits an unknown vulnerability (zero day).

Duqu, discovered on September 1, 2011 by CrySyS Lab,[12] is an infamous sophisticated cyberespionage malware thought to be related to the famous Stuxnet[13] 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 (Stewin & Bystrov, 2015), (Pathak & Nanded, 2016). 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.[14] 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.[15,16]

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 the *n-gram* feature extraction method and existing leading anti-virus engines.

SFEM is light, fast, and high performing, meeting the security needs of today's organizations which generate, process, transfer, receive, and analyze a massive amount of documents each day. Cloud services such as Microsoft Office 365 and Google Drive, which store or process documents, can integrate *SFEM* with machine learning in order to detect malicious documents.

Microsoft Word legacy binary documents (*.doc) are beyond the scope of this paper as they have not been used as the default format since Microsoft Office 2007. Moreover, the old *.doc format is not XML-based, and its structure substantially differs from the new XML-based format. Thus, the *.doc format requires a feature extraction methodology that is specifically tailored to its structure. We also focus on PCs, the platform that is most widely used by organizations and individuals. This paper's contributions are:

- Presenting a novel structural feature extraction methodology (*SFEM*) for XML-based Office documents.
- Presenting the use of machine learning algorithms, which have been trained on features extracted by *SFEM*, for the detection of malicious Microsoft Word XML-based documents (*.docx).
- Conducting experiments to determine the methods for best feature selection, feature representation, and top-feature selection in this context, and comparing *SFEM* with the *n*-gram feature extraction method and leading anti-virus engines.

[5] http://www.emc.com/domains/rsa/index.htm
[6] http://www.emc.com/careers/index.htm
[7] www.kaspersky.com
[8] https://www.metascan-online.com/blog/kaspersky-discovers-massive-cyber-attack-losses-could-reach-1-billion
[9] http://securelist.com/blog/research/65414/obfuscated-malicious-office-documents-adopted-by-cybercriminals-around-the-world/
[10] http://www.dailytech.com/Office+2010+to+Launch+Today+Microsoft+Owns+94+Percent+of+the+Market/article18360.htm
[11] https://www4.symantec.com/mktginfo/whitepaper/ISTR/21347932_GA-internet-security-threat-report-volume-20-2015-social_v2.pdf
[12] https://www.crysys.hu/
[13] http://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet
[14] http://www.darkreading.com/endpoint/with-$325-million-in-extorted-payments-cryptowall-3-highlights-ransomware-threat/d/d-id/1322899
[15] http://arstechnica.com/security/2016/02/locky-crypto-ransomware-rides-in-on-malicious-word-document-macro/
[16] http://thehackernews.com/2016/02/locky-ransomware-decrypt.html