DETECTING MALWARE IN WEB ADVERTISEMENTS

By

Emmanuel Akoch

Reg. Number: MAY16/COMP/007U

Department of Computer Science and Engineering

School of Computing and Engineering
Uganda Technology and Management University

Supervisor

Professor John Ngubiri

Uganda Technology and Management University

A Proposal Submitted to the School of Computing and Engineering in Partial Fulfillment of the Requirements for the Award of Masters in Computing (Computer Security Option) of Uganda Technology and Management University (UTAMU)

May 2017

1 Introduction

1.1 Background

Information and Communication Technology (ICT) is an essential part of our lives today and few can imagine living without it. ICT is becoming more powerful, more accessible, and more widespread. This technology is playing key role in enhancing competitiveness, enabling development, and bringing progress to all levels of society. For example, farmers in developing countries have benefited from new ICT services such as real-time information about commodity prices and weather, and from the ease of money transfers.

The effectiveness of governments has increased as a result of their ability to provide citizen-centric online services and to involve citizens in governance. ICT have become key enablers of business and employment creation, and of productivity growth.

Today computers and handheld mobile devices are considered one of the most important things a person or establishment can utilize. Many companies, individuals and businesses rely on ICT for fast communications, data processing and market intelligence. ICT is playing an integral role in every industry, helping companies improve business processes, achieve cost efficiencies, drive revenue growth and maintain competitive advantage in the marketplace.

Due to its intelligent, ubiquitous, and cost effective nature, the Internet as one of the ICT has become an essential component of communication across the globe, between marketers, advertisers and customers. More specifically, websites have become customers' first port of call for seeking information and eventually purchasing goods or services, whether online or offline. Consequently, businesses are increasing their presence on the Internet and improve their advertising practices so that current and prospective customers receive ads with content that is relevant and meaningful to them[1].

Advertising products and services through the Internet are more effective and efficient for surely it saves on a lot of time. Advertisements are immediately published which is not restricted by time or any geographical boundaries thus reaching global consumers at large. It has brought business to the finger tips by which a product can be viewed, liked, order placed and even paid for it online.

The intelligent nature of the Internet has also made advertising more targeted. There are certain programs like Google's AdWords and AdSense which match up the advertisers with the content that the targeted market concentrates on. This comprises of contextual ads on search engine results page, blogs, email marketing, advertising networks, online classified advertising and Social Network advertising. Internet has thus added new dimensions to advertising. It is a common belief that what sees more, sells more and the Internet has just made advertisements visible to the common eye more.

The complexity of the online advertising technology and the move of major businesses to advertise, market and sell their products online has made it easier for malicious actors to abuse the system. Since malicious ads do not require any type of user interactions in order to execute their payload, it has rendered malvertising a silent killer. The mere fact of browsing to a website that has malicious ads is enough to start the infection chain. This is a very big problem in the advertising network because most sites, if not all, do have ads.

According to [2], it was reported that the middle of 2015 was filled with accounts of malvertising affecting almost every segment of the ad-supported Internet. One of the possible explanation is that malvertising is simply an easier way to infect site visitors than spamming out links to infected websites. It's much easier for an attacker to try and compromise a popular site or seek to host malicious ads on popular, high-traffic websites because it means they don't need to consider the complex nuancing of social engineering, eliminating one or more step in the bad guys' "pipeline." It was also reported that Ad companies often don't request a lot of information from people submitting ads, making it easy for criminals to masquerade as legitimate businesses and upload malicious ads, which can appear on any number of sites.

In [3], reports that fake advertisements are here to stay, too, with an increasing number of ad networks that take a user's browsing session hostage, whether to deliver malware, scams, or endless surveys.

Businesses and individuals cannot avoid advertisements but it is very important that these businesses and individuals are not attacked by cybercriminals through advertisements. It is therefore very important to design means of avoiding attacks by malware in ads. Techniques for malware detection in advertisement network is still a grey area.

1.2 Statement of the Problem

Innovation is a key player to success of most businesses and individuals in this digital error. Doing something different, smarter or better that makes a positive difference in terms of value, quality or productivity greatly adds value to success of businesses. ICT has proven to be the technology in the world that has greatly influenced innovation. It has dramatically changed the lives of individuals and organizations. Currently, online shopping, digital marketing, social networking, digital communication and mobile cloud computing are the best examples of change which came through the wave of ICT.

Advertisers innovatively applied web advertisement for the marketing and sale of products and services. This is because the web has proven to be the best avenues for advertisers, linking directly to the world of business and serving the largest world population without border limitations.

Due to web advertisement popularity, cybercriminals implant malicious codes in the adverts with a target to infect the web users. Detection approaches for such malicious codes in adverts are currently insufficient.

1.3 Significance of the Study

1. The study will improve on the users and applications fare share of system resources. This is because the technique will identify some of the malware before the malware is executed by the users actions.

2. The study will also improve on the trust, privacy and security of electronic systems and its users. This technique will reduce on the lack of trust between customers and sellers, consumer privacy concerns and lack of security measures for Internet-based businesses. Therefore, similar levels of acceptance as traditional commerce will be achieved when trust, privacy and security becomes a built-in part of the electronic system.

1.4 General Objective

The objective of the study is to minimize user's vulnerability from attackers through designing a malicious ad detection technique.

1.5 Specific Objectives

1. To crawl 1500 ad impersonations from Alexa's top advertisement websites for analysis and manually verify normal legitimate ads and malicious ads using malformed iFrames, eval function, unescape function, shell codes, and whitespace randomization features

2. To analyze behavioral properties of advertisement websites. Dynamic properties such as redirects, registry changes, memory dumps, drive-by download and file changes are considered

3. To train a classifier basing on the features from static and dynamic analysis using K-NN algorithm

4. To test and validate the technique using K-Folds cross-validator

1.6 Scope of the Study

Two types of analysis will be applied i.e. static and dynamic. The proposed malware detection technique will be capable of detecting extracted features from the advertisements. The following features will be used as a basis to evaluate the technique:

Static analysis features:

- Malformed iFrames

- Whitespace randomization
- eval function
- unescape function
- shell codes

Behavioral analysis features:

- Drive-by Download
- File changes
- Memory Dumps
- Registry Changes
- Redirects

The system will not have a provision to clean/delete the identified suspicious ads from the advertisement network.

2 Literature Review

2.1 Advancement in Web Technology

Today, advances in Information and Communication Technology is leading in bringing together all categories of people both rich and poor, rural and urban, researchers and practitioners with the aim of working together towards a better life.

There have been many important advances in business technology this century, almost all of them enabled by ubiquitous broadband internet access, improved software development tools and the scalability and reliability of data centers.

Mobility is key to every business success. Everything from sales enablement, content marketing, and customer relations are handled through a click of a button. With the advancement in web technology, more people are using mobile devices to buy, sell, shop, find local businesses, and share their retail experiences with friends, acquaintances, prospects, and strangers every day.

With cloud computing on the other hand, businesses large and small have moved some of their operations to a third-party servers accessible over the internet connectivity. Not only does this allow for variable data packages but also for rapid (on-demand) expansion and mobility without the fear of downtime, crashes, or permanently lost data. This has allowed small business access to resources that would have been cost prohibitive for them in the past and evened the playing field when it comes to competing against corporations with far more funding.

Therefore, as different businesses compete with each other, the commercial advantage one can have over another may depend primarily on its use of information technologies. For example being able to extract information as to what the customer really wants and how to provide for that want can provide a significant advantage. This extraction of information is facilitated and indeed made possible by the technology used to store and manipulate this information. As the hardware and software mechanisms used to store and manipulate the information become more sophisticated and quicker, the business can utilize its stored information to maximize its commercial advantage.

2.2 Web Application Security

In today's Information and Communication Technology world, most businesses have a website that serves as the hub of their marketing ventures. It's the digital billboard and brochure that allows most businesses to reach their target audiences, but those websites are more susceptible to attacks.

In [4], it was reported that 86 percent of all websites contained at least one serious vulnerability, if not more. In that same year, Google reported that hacking had increased by 180% [5]. The web does not represent a threat for businesses, but cybercriminals use the web to scam legitimate businesses and individuals with aim of making profit, curiosity, self-expression or fun[6]. The attack are often initiated through spam email, internet advertisements, links in forums/social media, and fraudulent websites [7].

The current state of online advertising endangers the security and privacy of users. By tracking users on the Internet, advertisers can expose their personal activities and obtain information such as consulted web pages and social network connections [8]. The users can incur malware attacks through online ads without having to take any action other than visiting the mainstream website. According to [9], Yahoo confirms that users were served malicious advertisements, or "malvertisements", that if clicked, the advertisement directed users to websites that tried to install malicious software.

Some ad industry companies use automated systems to scan for malicious ads, but cybercriminals can learn the location of such scanners and not serve the ads to them. In other cases, attackers can change the content of a benign advertisement after it's been scanned and cleared [10].

The online advertising industry has grown in complexity to such an extent that each party can conceivably claim it is not responsible when malware is delivered to a user's computer through an advertisement. An ordinary online advertisement typically goes through five or six intermediaries before being delivered to a user's browser, and the ad networks themselves rarely deliver the actual advertisement from their own servers. In most cases, the owners of the host website visited by a user do not know what advertisements will be shown on their site [10].

This makes it impossible for ordinary consumers to avoid malicious ads, to identify their source and to determine whether the website they visited or the ad network could have prevented the attack. Therefore, loopholes in the currently applied advertising policies and the vulnerabilities that are exploited to attack users by serving malicious ads on user applications have been highlighted.

2.3 Malware Detection Techniques

A malware detection program D is the computational function that works in a domain which contains a collection of application programs $'P'$, and a collection of malicious and benign programs. The detector program $'D'$ analyzes the programs $'p'$ which belongs to the set of application programs $'P'$ to find whether it is a benign (normal program), or a malware (malicious program) [11]. The detection program determines the identity of a program by analysis or by identification. But sometimes this function may result in, false positive, false negative or undecidable objects depending on efficiency of the function $'D'$.

Undecidability is for zero day malware (new unknown malware), as classification methods fail to determine the identity of a program. False positive is a malware detected while it is not malware and false negative is a benign program detected while it is not benign [12].

According to [13], malware detector is the implementation of some malware detection technique(s). The malware detector attempts to help protect the system by detecting malicious behavior. The malware detector may or may not reside on the same system it is trying to protect. The malware detector performs its protection through the manifested malware detection technique(s), and serves as an empirical means of evaluating malware detection techniques' detection capabilities.

Malware detectors take two inputs. One input is its knowledge of the malicious behavior. In anomaly-based detection, the inverse of this knowledge comes from the learning phase. So theoretically, anomaly-based detection knows what is anomalous behavior based on its knowledge of what is normal. Since anomalous behavior subsumes malicious behavior, some sense of maliciousness is captured by anomaly based detection.

The other input that the malware detector must take as input is the program under inspection. Once the malware detector has the knowledge of what is considered malicious behavior (normal behavior) and the program under inspection, it can employ its detection technique to decide if the program is malicious or benign.

Techniques used for detecting malware can be categorized broadly into two categories: anomaly based detection and signature-based detection. This research proposal emphasizes anomaly based detection technique.

2.4 Anomaly-Based Detection Approach

An anomaly-based detection technique uses its knowledge of what constitutes normal behavior to decide the maliciousness of a program under inspection [13]. The main purpose of anomaly based detection approach is to analyze the behavior of known or unknown malware [14] [15]. Anomaly-based detection usually occurs in two phases–a training (learning) phase and a detection (monitoring) phase [13] [14]. During training phase the behavior of system is observed in the absence of attack and machine. During the training phase the detector attempts to learn the normal behavior. The detector could be learning the behavior of the host or the PUI (Physical User Interface) or a combination of both during the training phase. A key advantage of anomaly-based detection is its ability to detect zero-day attacks. Similar to zero-day exploits, zero-day attacks are attacks that are previously unknown to the malware detector [13]. The two fundamental limitations of this technique is its high false alarm rate and the complexity involved in determining what features should be learned in the training phase [13][11][16][15][17].

2.5 Anomaly-Based Detection Algorithm

Support Vector Machine (SVM)

A Support Vector Machine performs classification by constructing an N dimensional hyper plane that optimally separates the data into two categories [18].

In the parlance of SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyper plane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyper plane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane. The vectors near the hyper plane are the support vectors.

Researchers [19] applied anomaly based detection technique to identify malicious advertisements at the publishers' end. They based they framework on two types of analysis of which one included the behavioral analysis of the advertisements done in a secure sandboxed environment to detect malicious activities. They extracted a total of 9 features from 1500 advertisements and classified it using a trained one class SVM classifier. Their result shows that 53% of the suspicious ads contain dubious iFrames while 69% of them perform redirections followed by drive by download 18% with very low false positive and false negative rates.

K-Nearest Neighbor (K-NN)

K-Nearest Neighbors (K-NN) is one of the simplest, though, accurate machine learning algorithms. K-NN is a non-parametric algorithm, meaning that it does not make any assumptions about the data structure. In real world problems, data rarely obeys the general theoretical assumptions, making non-parametric algorithms a good solution for such problems. K-NN model representation is as simple as the dataset – there is no learning required, the entire training set is stored.

The K-nearest neighbor algorithm is a method for classifying objects based on closest training examples in the feature space.

K-NN classification divides data into a test set and a training set. For each row of the test set, the K nearest (in Euclidean distance) training set objects are found, and the classification is determined by majority vote with ties broken at random. If there are ties for the Kth nearest vector, all candidates are included in the vote.

The way in which the algorithm decides which of the points from the training set are similar enough to be considered when choosing the class to predict for a new observation is to pick the k closest data points to the new observation, and to take the most common class among these [20].

With K-NN algorithm, all the three distance measures (Euclidean, Manhattan and Minkowski) are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset. The algorithm (as described in [20]) can be summarized as:

a) A positive integer k is specified, along with a new sample

b) We select the k entries in our database which are closest to the new sample

c) We find the most common classification of these entries

d) This is the classification we give to the new sample

According to [21], it has been possible to extract 24 malware candidates out of 2441 original candidates from which 25% are surely malicious and 50% which are probably malicious, have to be further investigated in order to obtain a decisive classification.

Researchers [22] noted that the performance of the K-NN classifier algorithm also depends on the value of k, the number of nearest neighbors of the test process. Usually the optimal value of k is empirically determined. They also noted that the K-NN classifier doesn't have to build separate profiles of short system call sequences for different programs, thus the calculations involved with classifying new program behavior is largely reduced.

K-NN classifier works well with dynamic environments that requires frequent updates of the training data, which makes it attractive for intrusion detection [22] as well as malicious code detection.

Decision Trees

A decision tree is a hierarchical data structure implementing the divide-and-conquer strategy. It is an efficient nonparametric method, which can be used for both classification and regression [23].

A decision tree is a predictor, $h : x \longrightarrow y$ that predicts the label associated with an instance x by traveling from a root node of a tree to a leaf. For simplicity we focus on the binary classification setting, namely, $y = \{0, 1\}$, but decision trees can be applied for other prediction problems as well.

At each node on the root-to-leaf path, the successor child is chosen on the basis of a splitting of the input space. Usually, the splitting is based on one of the features of x or on a predefined set of splitting rules. A leaf contains a specific label [24].

3 Methodology

3.1 Static Analysis

This includes crawling 1500 ad impersonations from Alexa's top advertisement websites. The source code of the advertisement will be analyzed basing on the below features:

- **Malformed iFrames:** This is used to automatically redirect users to unintended pages. Attackers can easily embed hidden iFrames that serve malvertisements while interacting with a legitimate user [19].Therefore, the aim of this static code analysis is to locate iFrames with width and height properties values set to negative. Also, if the size of iFrame is small (e.g. <iframe width="1" height="1" frameborder="0" scrolling="no" marginheight="0" marginwidth="0" src="http://badnetwork.com/getbadfile.php"> </iframe>) or null, the framework will classify that iFrame as suspicious. For large iFrames, the presence of object tag is checked. Since attackers use Object tags to embed malicious scripts, the presence of object tag is classified as suspicious.

- **Whitespace randomization and *eval* Function:** The JavaScript *eval* function evaluates or executes an argument. The presence of eval function is checked. Self-including DOM (Document Object Model) worker XSS (Cross-site Scripting) and hidden *eval()* are the features that will be inspected in this case. The other analysis in the static module includes whitespace randomization checking.

- ***unescape* function:** Shell codes obfuscation techniques using *unescape* funcation has been used to distribute malware[19]. Despite the *unescape* feature being removed from the Web standards, some browsers may still support it. Therefore, the presence of escaped characters and *unescape* functions in an advertisement will be classified as suspicious during this static code analysis process.

- **Shell codes:** Shell codes obfuscation techniques using *unescape* has been used to distribute malware. The presence of JavaScript unescape function with large amount of escaped data could be suspected to potentially house large amount of shell code or malicious JavaScript. The presence of any VB script or shell code that can be used as an anti-detection mechanism by attackers will also be checked.

3.2 Dynamic Analysis

The dynamic analysis will be used to check the behavior of the advertisements in a sandboxed environment. In the behavior analysis, the dynamic properties of the Operating Systems such as registry changes, file changes and memory dumps will be checked. The network will also be monitored for changes such as redirects and drive-by download. These properties will then be considered as a second set of features to be examined.

Cuckoo Sandbox, a free software that automates the task of analyzing any malicious file under Windows, OS X, Linux and Android will be used for dynamic analysis. Cuckoo sandbox will be used because it can analyze many different malicious files (executable, document exploits, Java applets) as well as malicious websites, in Windows, OS X, Linux, and Android virtualized environments. Trace API (Application Program Interface) calls and general behavior of the file. Dump and analyze network traffic, even when encrypted. Perform advanced memory analysis of the infected virtualized system with integrated support for Volatility. The web pages with the ads will be launched in multiple web browsers in a Cuckoo sandbox environment. Internet explorer, Google Chrome and Mozilla Firefox web browsers will be used. This is because they are some of the most popular web browsers. The idea of the behavior analysis is that an advertisement is not supposed to make file changes or spawn processes or even make suspicious redirects. Hence any such events are taken as a suspicious feature[19].

3.3 Classification using K-NN Algorithm

The training phase of the algorithm will consist of storing the feature vector X and the class label C of the training sample. Each element x_i of the feature set will be used as inputs to the trained K-NN classifier. The feature set elements x_i represents a Boolean value which indicates whether an attack method found in the static analysis is present or absent.

In the classification phase, k will be a user-defined constant which will be used for searching through the entire training set for the most k similar instances (the neighbors) and summarizing the output variable for those k instances. Hamming distance metric

$D_H = \sum_{i=1}^{k} |x_i - y_i|$ will be used for distance measurement function. The classifier will be trained on the dataset using scikit-learn [25].

3.4 To test and validate the technique

K-Fold cross-validation will be used to validate the technique. The training dataset will be partitioned into two pieces: training and test, where K represents the number of folds or observations to take place. The number of folds will depend on the value of k. $KFolds$ divides all the samples in k groups of samples, called folds (if $k = n$, this is equivalent to the Leave One Out strategy), of equal sizes (if possible). The prediction function is learned using $k - 1$ folds, and the fold left out is used for test [26].

The accuracy of the algorithm will also be evaluated. The accuracy (ACC) is defined as the number of correctly classified events over the total number of events using:

$$ACC = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

Where;

- True Positive (TP) = Number of samples correctly predicted as malware
- False Positive (FP) = Number of samples incorrectly predicted as malware
- True Negative (TN) = Number of samples correctly predicted as benign
- False Negative (FN) = Number of samples incorrectly predicted as benign.

References

[1] Eurostat, Internet advertising of businesses - Statistics on Usage of Ads, December 2016

[2] Symantec, “Internet Security Threat Report,” Symantec Corporation, 2016.

[3] McAfee. Labs, “2017 Threats Predictions,” Intel Security, 2016

[4] WhiteHat Security, Website Security Statistics Report, 2015

[5] Google Webmaster Central Blog, Helping hacked sites reconsideration requests. September 2015. https://webmasters.googleblog.com/2015/09/helping-hacked- sites-with.html

[6] Kaspersky Lab. Daily. What Motivates Cybercriminals? Money, Of Course. https://blog.kaspersky.com/what-motivates-cybercriminals- money-of-course/717/

[7] Federal Bureau of Investigation. Internet Crime Complaint Center (IC3). 2016 Internet Crime Report

[8] Julient Freudiger, Nevena Vratonjic and Jean-Pierre Hubaux, Towards Privacy-Friendly Online Advertising

[9] http://www.pcworld.com/article/2086700/yahoo- malvertising-attack-linked-to-larger-malware-scheme.html

[10] United States Senate. Permanet Subcommittee on Investigations.Committee on Homeland Security and Governmental Affairs. Online Advertising and Hidden Hazards to Consumer Security and Data Privacy, May 15, 2014 Hearing

[11] Imtithal A. Saeed, Ali Selamat, and Ali M. A. Abuagoub, A survey on malware and malware detection systems, International Journal of Computer Applications, vol. 67, p. 0975 - 8887, April, 2013.

[12] Malwarebytes Labs, “State of Malware Report, 2017,” 2017.

[13] N. Idika and A. P. Mathur, A survey of malware detection techniques, no. 47907.

[14] J. Landage and M. P.Wankhade, Malware and malware detection techniques: A survey, International Journal of Engineering Research & Technology (IJERT), no. IJERTIJERT ISSN: 2278-0181, December, 2013.

[15] A. Mujumdar, G. Masiwal and B. B. Meshram, Analysis of signaturebased and behavior-based anti-malware approaches, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), no. ISSN: 2278 - 1323, June, 2013.

[16] A. S. Bist. et al., Classication and identication of malicious codes, Indian Journal of Computer Science and Engineering (IJCSE), vol. 3, no. ISSN : 0976-5166, May, 2012.

[17] SANS Institute InfoSec Reading Room, “The Expanding Role of Data Analytics in Threat Detection,” SANS Institute, 2015.

[18] T. O. Ayodele, Types of Machine Learning Algorithms. University of Portsmouth United Kingdom.

[19] P. Poornachandran et al., Demalvertising: A Kernel Approach for Detecting Malwares in Advertising Networks, Amrita Center for Cyber Security Systems and Networks Amrita Vishwa Vidyapeetham Amritapuri Campus, Kollam, India, 2017.

[20] O. Sutton, Introduction to k nearest neighbour classication and condensed nearest neighbour data reduction, February, 2012.

[21] J. Hegedus, Y. Miche, A. Ilin and A. Lendasse, Methodology for Behavioral based Malware Analysis and Detection using random projections and K-nearest neighbors classiers, Department of Information and Computer Science, Aalto University School of Science, FI-00076 Aalto, Finland.

[22] Y. Liao and V. R. Vemuri, Using K-Nearest Neighbor Classier for Intrusion Detection.

[23] E. Alpaydin, Introduction to Machine Learning, Second Edition. Massachusetts Institute of Technology, 2010. ISBN 978-0-262-01243-0

[24] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. [Online]. Available: http://www.cs.huji.ac.il/ shais/UnderstandingMachineLearning

[25] http://scikitlearn.org/stable/modules/generated/sklearn.neighbors

[26] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html