Stegcheck and URL detection using machine learning

Mrs. C P Jetlin[1], Bansie V Iyengar[2], Kamalesh Hari M U[3], Gopi M[4]
[1]*M.E., Ph.D., Asst. Professor, Department of CSE, Agni College of Technology, Chennai*
[2,3,4]*UG Student, Department of CSE, Agni College of Technology, Chennai*

Abstract - **This paper presents three security concepts. The first two approaches are of the form of image and audio steganography and the third security concept that is covered is malicious URL detection. Steganography is the technique of hiding secret data in order to avoid detection, the secret data is then extracted at its destination. The use of steganography can be combined with encryption as an extra step for hiding or protecting data. Depending on the nature of the cover object, steganography can be divided into four types: Text Steganography, Image Steganography, Video Steganography, Audio Steganography. Steganography is a means of storing data in a way that it hides the existence of them. using steganography to communicate greatly reduces the risk of information leakage. steganography enhances the privacy individually, although it is not a substitute for encryption. Malicious Web sites are a cornerstone of Internet criminal activities. These Web sites contain various unwanted content such as spam-advertised products, phishing sites, dangerous "drive-by" harness that infect a visitor's system with malware. Despite a growing number of vendors offering anti-phishing solutions, phishing is a bigger problem than ever. The problem is so big, in fact, that it is hard to keep up with the latest facts and figures. The average breach costs organizations $3.92 million. This number will generally be higher in larger organizations and lower in smaller organizations. While the Manufacturing industry saw the most breaches from social attacks, employees working in Wholesale Trade are the most frequently targeted by phishing attacks, with 1 in every 22 users being targeted by a phishing email last year. In the proposed system we are using Machine-Learning techniques to classify a URL as either safe or unsafe in Real Time without even the need to download the webpage. We will be using the Logistic regression model, where the Labelled URLs will be pre-processed by eliminating extras like the 3rd level domains. From this project, we hope to build alternative security solution for the above discussed problems.**

I.INTRODUCTION

Internet has emerged as the most convenient and efficient medium for communication. Through internet, messages can be transferred in a fast and cheap way in various fields like government offices, private sector, military, and medical areas. Many times, confidentiality of the transferred message needs to be maintained. To ensure that the message is transferred securely and safely over the network, a suitable method is needed. Steganography proves as a trustable method for achieving this aim. In steganography, the data are hidden in the cover media. The cover medium can be in the form of image file, text file, video file, or audio file. The most popular medium used is image files because of their high capacity and easy availability over the internet. At the sender's side, the image used for embedding the secret message is called cover image, and the secret information that needs to be protected is called a message. As soon as data are embedded using some appropriate embedding algorithm, then it is called stego image. This stego image is transferred to the receiver, and he extracts out the secret message using extraction algorithm. Various features influence the quality of audio steganographic methods. The importance and the impact of each feature depend on the application and the transmission environment. The most important properties include robustness to noise, to compression and to signal manipulation, as well as the security and the hiding-capacity of hidden data. The robustness requirement is tightly coupled with the application and is also the most challenging requirement to fulfill in a steganographic system when traded with data hiding-capacity. The importance of the World Wide Web has continuously been increasing. Unfortunately, the technological advancements come coupled with new sophisticated techniques to attack and scam users. Such attacks include rogue websites that sell counterfeit goods, financial fraud by tricking users into revealing

sensitive information which eventually lead to theft of money or identity, or even installing malware in the user's system. There are a wide variety of techniques to implement such attacks, such as explicit hacking attempts, drive-by download, social engineering, phishing, watering hole, man-in-the middle, SQL injections, loss/theft of devices, denial of service, distributed denial of service, and many others. Considering the variety of attacks, potentially new attack types, and the innumerable contexts in which such attacks can appear, it is hard to design robust systems to detect cyber-security breaches. The limitations of traditional security management technologies are becoming more and more serious given this exponential growth of new security threats, rapid changes of new IT technologies, and significant shortage of security professionals. Most of these attacking techniques are realized through spreading compromised URLs (or the spreading of such URLs forms a critical part of the attacking operation).

II.RELATED WORK

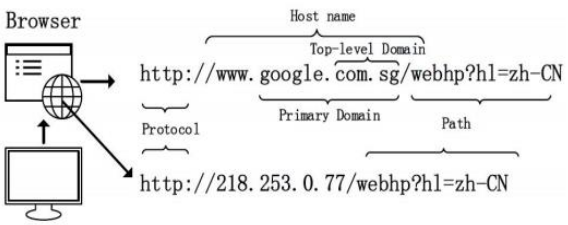
Literature Survey

Steganography has emerged as a glowing research area in which various methods have been proposed in several carrier media. Here, we are going to present the brief overview of some already proposed methods, and at last, we will be presenting the comparison work with these methods.

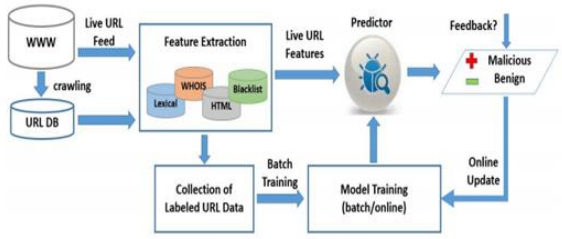
LSB method provides the very basic idea of steganography in an easy manner. This method states that the secret message bits can be placed by replacing the least significant bits of the pixels of the image. It allows 100% insertion of message binary bits in the pixels of an image with a very minute change of +1 or −1 in the value of the pixels. This method was vulnerable to attack as the message was present at LSB, and by only picking LSBs, the intruder can access the data. Quantization noise can also destroy the data present on LSB. So, this method can be easily decoded by the intruder and is also not immune to the noise and compression techniques. Also, this method allows only single bit insertion of message data inside the particular pixel. This method was vulnerable to attack as the message was present at LSB, and by only picking LSBs, the intruder can access the data. Singh et al. proposed a method based on first and second bit plane. In this method, on the combination of 1st and 2nd bit plane the message was hidden. The main result of the method was that the probability of message insertion at a pseudorandom location at first chance was 50%. The probability was 50% when there was no need to change the pixel value. The probability was 12.5% when a change in pixel value was required. Batra and Rishi proposed a method in which the message was hidden using the 6th, 7th, and 8th bits of a pixel in a grayscale image. This method overcomes the limitation of the Singh et al.'s method. The main result of the method was that the probability of message insertion at a pseudorandom location at first chance is 85.93%. The probability when the message was not changed was 43.18%. As the result shows, this method does not provide 100% message insertion rate. In FMM (Five Modulus method) , the cover image was divided into N blocks with block size k∗k pixels where k is the size of the window. Each pixel in these blocks was modified such that the pixel of the block is divisible by 5. The beauty of this method was that the message was scattered over the entire image. The limitation of this method is the hiding capacity that is low. In some cases, the average message hidden capacity is below 1 bit per pixel. Bailey and Curran have presented the facts of the Stego Color Cycle (SCC) method. This is the advancing method of LSB. Here also LSB of pixels of color images is used for insertion of secret message binary bits. Insertion is done in a cyclic way by choosing the LSB of the red channel of first pixel and then LSB of the green channel of second pixel and then LSB of the blue channel of third pixel, and this cycle repeats in same cyclic order for all the pixels. This method also allows 100% insertion for RGB images, but for its simple cyclic order, it can be easily decoded by the intruder.

III.METHODOLOGY

URL is the abbreviation of Uniform Resource Locator, which is the global address of documents and other resources on the World Wide Web. A URL has two main components: (i) protocol identifier (indicates what protocol to use) (ii) resource name (specifies the IP address or the domain name where the resource is located). The protocol identifier and the resource name are separated by a colon and two forward slashes.



A variety of approaches have been attempted to tackle the problem of Malicious URL Detection. According to the fundamental principles, we categorize them into: (i) Blacklisting or Heuristics, and (ii) Machine Learning approaches.



IV. DATA PROCUREMENT

For utilizing the application, the driver plays out an enrollment if utilizing the application interestingly. Subsequent to playing out a sign-up, the driver adds a ride by entering the source and objective of the ride. Similarly, an interface for the travelers is additionally furnished where the travelers can associate with the ride, added by the driver. The driver at that point begins the ride. The proposed application at that point catches the continuous pictures of the driver. Pictures are caught each time the application gets a reaction from the worker. The cycle goes on until the driver stops the ride. For testing the effectiveness of the proposed approach, an informational index of 50 volunteers was gathered. Each member was approached to flicker their eyes irregularly while taking a gander at the camera for catching EAR esteems. The logs of the outcomes that were caught by the application were gathered and broke down with the assistance of AI classifiers.

V. PROBLEM FORMULATION

We formulate Malicious URL detection as a binary classification task for two-class prediction: “malicious" versus “benign". Specifically, given a data set withT URLs {(u1,y1), . . . ,(uT ,yT)}, where ut for t = 1, . . . ,T represents a URL from the training data, and yt ∈ {1, −1} is the corresponding label where yt = 1 represents a malicious URL and yt = −1 represents a benign URL. The crux to automated malicious URL detection is two-fold: (1) Feature Representation: Extracting the appropriate feature representation: ut → xt where xt ∈ R d is a d-dimensional feature vector representing the URL; and (2) Machine Learning: Learning a prediction function f : R d → R which predicts the class assignment for any URL instance x using proper feature representations. The goal of machine learning for malicious URL detection is to maximize the predictive accuracy. Both of the folds above are important to achieve this goal. While the first part of feature representation is often based on domain knowledge and heuristics, the second part focuses on training the classification model via a data driven optimization approach. Fig. 2 illustrates a general workflow for Malicious URL Detection using machine learning. The first key step is to convert a URL u into a feature vector x, where several types of information can be considered, and different techniques can be used. Unlike learning the prediction model, this part cannot be directly computed by a mathematical function (not for most of it). Using domain knowledge and related expertise, a feature representation is constructed by crawling all relevant information about the URL. These range from lexical information (length of URL, the words used in the URL, etc.) to host-based information (WHOIS info, IP address, location, etc.). Once the information is gathered, it is processed to be stored in a feature vector x. Numerical features can be stored in x as is, and identity related information or lexical features are usually stored through a binarization or bag-of-words (BoW) approach. Based on the type of information used, x ∈ R d generated from a URL is a d-dimensional vector where d can be less than 100 or can be in the order of millions. A unique challenge that affects this problem setting is that the number of features may not be fixed or known in advance. For example, using a BoW approach one can track the occurrence for every type of word that may have occurred in a URL in the training data. A model can be trained on this data, but while predicting, new URLs may have words that did not occur in the training data. It is thus a challenging task to design a good feature representation that is robust to unseen data. After obtaining the feature vector x for the training data, to learn the prediction function f: R d →

R, it is usually formulated as an optimization problem such that the detection accuracy is maximized (or alternately, a loss function is minimized). The function f is (usually) parameterized by a d− dimensional weight vector w, such that f(x) = (w⊤x). Let yˆt = sign (f(xt)) denote the class label prediction made by the function f. The number of mistakes made by the prediction model on the entire training data is given by: ÍT t=1 Iyˆt =yt where I is an indicator which evaluates to 1 if the, Vol. 1, No. 1, Article. Publication date: August 2019. Malicious URL Detection using Machine Learning: A Survey 7 condition is true, and 0 otherwise. Since the indicator function is not convex, the optimization can be difficult to solve. As a result, a convex loss function is often defined, and is denoted by ℓ(f(x),y) and the entire optimization can be formulated as: min w Õ T t=1 ℓ(f(xt),yt) (1) Several types of loss functions can be used, including the popular hinge-loss ℓ(f(x),y) = 1 2 max(1− yf(x), 0), or the squared-loss ℓ(f(x),y) = 1 2 (f(x) − y) 2 . Sometimes, a regularization term is added to prevent over-fitting or to learn sparse models, or the loss function is modified based on cost sensitivity of the data (e.g., imbalanced class distribution, different costs for diverse threats).

Feature Representation:

The process of feature representation can be further broken down into two steps: (1) Feature Collection: This phase is engineering oriented, which aims to collect relevant information about the URL. This includes information such as presence of the URLs in a blacklist, features obtained from the URL String, information about the host, the content of the website such as HTML and JavaScript, popularity information, etc. Figure 3 gives an example to demonstrate various types of various types of information that can be collected from a URL to obtain the feature representation. (2) Feature Preprocessing: In this phase, the unstructured information about the URL (e.g., textual description) is appropriately formatted, and converted to a numerical vector so that it can be, Vol. 1, No. 1, Article. Publication date: August 2019. 8 Sahoo, Liu and Hoi fed into machine learning algorithms. For example, the numerical information can be used as is, and Bag-of-words is often used for representing textual or lexical content. For malicious URL detection, researchers have proposed several types of features that can be used to provide useful information. We categorize these features into: Blacklist Features, URL-based Lexical Features, Host-based features, Content-based Features, and Others (Context and Popularity). All have their benefits and short-comings - while some are very informative, obtaining these features can be very expensive. Similarly, different features have different preprocessing challenges and security concerns.

Hiding in temporal domain

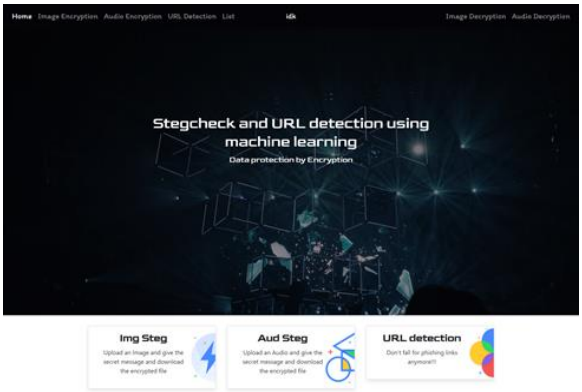
Low bit encoding

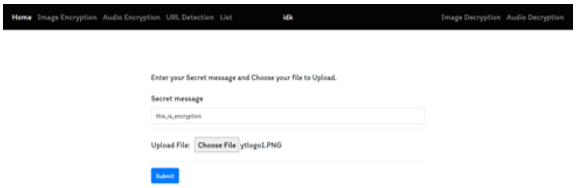
Also known as LSB (Least Significant Bit), this method is one of the earliest methods used for information hiding. Traditionally, It is based on embedding each bit from the message in the least significant bit of the cover audio in a deterministic way.

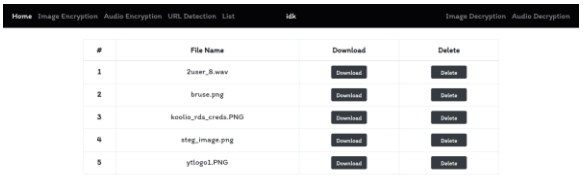
Hiding in silence intervals:

In, a simple and effective embedding method has been used to exploit silence intervals in speech signal. Initially, the silence intervals of the speech and their respective lengths (the number of samples in a silence interval) are determined. These values are decreased by a value x where $0 < x < 2^{nbits}$, and nbits is the number of bits needed to represent a value from the message to hide.

VI.PERFORMANCE EVALUATION WITH EXPERIMENTAL RESULTS AND DISCUSSION







VII. CONCLUSION AND FUTURE WORK

Malicious URL detection plays a critical role for many cybersecurity applications, and clearly machine learning approaches are a promising direction. In this article, we conducted a comprehensive and systematic survey on Malicious URL Detection using machine learning techniques. In particular, we offered a systematic formulation of Malicious URL detection from a machine learning perspective, and then detailed the discussions of existing studies for malicious URL detection, particularly in the forms of developing new feature representations, and designing new learning algorithms for resolving the malicious URL detection tasks. In this survey, we categorized most, if not all, the existing contributions for malicious URL detection in literature, and also identified the requirements and challenges for developing Malicious URL Detection as a service for real-world cybersecurity applications.
In order to provide better protection to digital data content, new steganography techniques have been investigated in recent researcher works. The availability and popularity of digital audio signals have made them an appealing choice to convey secret information. Audio steganography techniques address issues related to the need to secure and preserve the integrity of data hidden in voice communications in particular. In an attempt to reveal their capabilities in ensuring secure communications, we discussed their strengths and weaknesses.

REFERENCES

[1] Neda Abdelhamid, Aladdin Ayesh, and Fadi Thabtah. 2014. Phishing detection based associative classification data mining. Expert Systems with Applications (2014).
[2] Farhan Douksieh Abdi and Lian Wenjuan. 2017. Malicious URL Detection using Convolutional Neural Network. Journal International Journal of Computer Science, Engineering, and Information Technology (2017).
[3] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. [n. d.]. A comparison of machine learning techniques for phishing detection. In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit. , Vol. 1, No. 1, Article . Publication date: August 2019. Malicious URL Detection using Machine Learning: A Survey 29 ACM.
[4] Sadia Afroz and Rachel Greenstadt. 2011. Phishzoo: Detecting phishing websites by looking at them. In Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. IEEE.
[5] Anupama Aggarwal, Ashwin Rajadesingan, and Ponnurangam Kumaraguru. 2012. Phishari: automatic realtime phishing detection on twitter. In eCrime Researchers Summit (eCrime), 2012. IEEE.
[6] Djebbar F, Guerchi D, Abed-Maraim K, Hamam H: Text hiding in high frequency components of speech spectrum, Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on, vol., no., (Malaysia, 10-13 May 2010).
[7] Shirali-Shahreza S, Shirali-Shahreza M: Steganography in Silence Intervals of Speech, proceedings of the Fourth IEEE International Conference on Intelligent Information Hiding and Multimedia Signal (IIH-MSP 2008). (Harbin, China, August 15-17, 2008).