LITERATURE SURVEY ON WEB PHISING DETECTION

ABSTRACT:

This article surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber-attacks are spread via mechanisms that exploit weaknesses found in end-users, which makes users the weakest element in the security chain. This paper aims at surveying many of the recently proposed phishing mitigation techniques. A high-level overview of various categories of phishing mitigation techniques is also presented, such as: detection, offensive defences, correction, and prevention, which we belief is critical to present where the phishing detection techniques fit in the overall mitigation process.

INTRODUCTION:

Phishing is a social engineering attack that aims at exploiting the weakness found in system processes as caused by system users. Phishing is a type of attack that used to induce online user to get their personal information such as credit card number, bank details which causes financial loss of user. In last few years the use of online social network becomes more popular such as You Tube, Facebook and Twitter. All over world people use SNS to communicate with other people they share their data with them. The shared data may be confidential. As a use of SNS increases parallel the creation of computer malware, such as viruses and phishing attacks, has also continued to rise. One of the most famous examples of this new attack is the Koobface virus. The Koobface virus spreads through hyperlinks that appear to come from one of your friends, usually advertising a funny video. When the victim clicks the link to watch the video. They are met with a pop-up message stating that they need to update their Adobe Flash player. When the user clicks to download the “update”, they are actually downloading a Trojan horse which installs both a web proxy and a backdoor on the victim’s system.

According to RSA report in year 2014 nearly RSA identified 33, 145 phishing attacks in August, RSA estimates phishing cost global organizations $282 million in losses. The U.S. remained the most targeted country in August with 61% of phishing volume. China, Netherlands, United Kingdom and Canada were collectively targeted by 20% of total attacks. Hong Kong remained the second top hosting country for phishing in August with 13% of total attacks. The volume of attacks hosted in Italy doubled from July from 3% to 6% [1].

According to APWG global phishing report there were at least 123, 741 unique phishing attacks worldwide. Most of the growth in attacks came from increases in attacks against vulnerable hosting and use of maliciously registered domains and subdomains. An attackis defined as a phishing site that targets a specific brand. The number of domain names in the world grew from 271.5 million in November 2013 to 279.5 million in April 2014.It identifies 22, 679 domain names registered maliciously by phishers and most of these registrations were made by Chinese phishers, especially using free domain name registrations. The targets included more large and small banks in Latin America, India, and the Middle East. The phishing attack using social networking site and email is around 23.1%, 25.7%, 32.4% and 12.8% for bank, ecommerce industry and money transaction respectively over total phishing attack [2].

RELATED WORK:

Nowadays most people use internet for various purposes such as online shopping like purchasing or selling products, chat with friends, sending mail. Internet users now spend more time on social networking sites Information can spread very fast and easily within the social media networks. Social media systems depend on users for content contribution and sharing. Facebook had over 1.3 billion active users as of June 2014. there are over 1.3 billion (the number is keep growing) pages from various categories, such as company, product/service, musician/band, local business, politician, government, actor/director, artist, athlete, author, book, health, beauty, movie, cars, clothing, community. Fans not only can see information submitted by the page, but also can post comments, photos and videos to the page.

Justin Ma et. al proposed Identifying Suspicious URLs: An Application of Large-Scale Online Learning in which explores online learning approaches for detecting malicious web sites using lexical and host-based features of the associated URLs. Use various lexical and host-based features of the URL for classification, but exclude web page content. If properly automated, this technique can afford the low classification overhead of blacklisting while offering far greater accuracy. in this proposed system built a URL classification system that uses a live feed of labelled URLs from a large web mail provider, and that collects features for the URLs in real-time [3].

Xin Jin et. al proposed Social Spam guard as spam detection System for Social Media Networks security. Due to the huge number of posts (over billions) on social media, manually

checking every post to pick up the spams is impossible. scalable active learning approach proposed to manually verify as many spams as possible This system has several benefits automatically harvesting spam activities in social network, Introducing both image and text content features and social network features to indicate spam activities Integrating with our GAD clustering algorithm to handle large scale data and Introducing a scalable active learning approach to identify existing spams with limited human efforts, and perform online active learning to detect spams in real-time[4].

DATA MINING:

Data mining is the process for extracting hidden knowledge and pattern from large database. Knowledge Discovery from Data is synonym for data mining. Machine learning approaches are commonly used to classify malicious URLs and anomalies. Machine learning uses different classification technique such as logistic regression, SVM, and Bayesian classification. Logistic regression is used to predict the outcome of a binary dependent variable based on various predictor variables. SVM represents training data as points in space and locates one hyperplane in order to classify the data into categories.

The point representing training data are support vectors and the solid line represents the separating hyperplane used for classifying the test data. Bayesian classification is statistical classifier.

It is based on Bayes theorem P $(H\ X)=P\ X/H\ P(H)\ p(X)$ X is a data tuple.

In Bayesian terms, X is considered evidence.

H be some hypothesis, such as that the data tuple

X belongs to a specified class C.

For classification problems, we want to Determine P(H / X), the probability that the hypothesis H holds given the “evidence” or observed data tuple X.

P (H), which is independent of X. P(X/H) is the posterior probability of X conditioned on H. P (H) is the prior probability, or a priori probability, of H.

P(X) is the prior probability of X.

CONCLUSION:

User education or training is an attempt to increase the technical awareness level of users to reduce their susceptibility to phishing attacks. It is generally assumed that the addition of user education materials compliments technical solutions (e.g., classifiers). However, the human factor is broad and education alone may not guarantee a positive behavioural response.