WEB PHISHING DETECTION

LITERATURE REVIEW

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. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks.

INTRODUCTION:

PHISHING is a social engineering attack that aims at exploiting the weakness found in system processes as caused by system users. For example, a system can be technically secure enough against password theft, however unaware end users may leak their passwords if an attacker asked them to update their passwords via a given Hypertext Transfer Protocol (HTTP) link, which ultimately threatens the overall security of the system.

Moreover, technical vulnerabilities (e.g., Domain Name System (DNS) cache poisoning) can be used by attackers to construct far more persuading socially-engineered messages (i.e., use of legitimate, but spoofed, domain names can be far more persuading than using different domain names). This makes phishing attacks a layered problem, and an effective mitigation would require addressing issues at the technical and human layers. Since phishing attacks aim at exploiting weaknesses found in humans (i.e., system end-users), it is difficult to mitigate them. For example, as evaluated in, end-users failed to detect 29% of phishing attacks even when trained with the best performing user awareness program. On the other hand, software phishing detection techniques are evaluated against bulk phishing attacks, which makes their performance practically unknown with regards to targeted forms of phishing attacks.

MAIN BODY:

A. HISTORY:

According to APWG, the term *phishing* was coined in 1996 due to social engineering attacks against America On-line (AOL) accounts by online scammers. Phishing attacks were historically started by stealing AOL accounts, and over the years moved into attacking more profitable targets, such as on-line banking and e-commerce services. Currently, phishing attacks do not only target system end-users, but also technical employees at service providers, and may deploy sophisticated techniques such as MITB attacks.

B. IMPORTANCE:

According to APWG, phishing attacks were in a raise till August, 2009 when the all-time high of 40,621 unique phishing reports were submitted to APWG. The total number of submitted unique phishing websites that were associated with the 40,621 submitted reports in August, 2009 was 56,362. Minimizing the impact of phishing attacks is extremely important and adds great value to the overall security of an organization.

C. CHALLENGES:

Because the phishing problem takes advantage of human ignorance or naivety with regards to their interaction with electronic communication channels (e.g., E-Mail, HTTP, etc...), It is not an easy problem to permanently solve. All of the proposed solutions attempt to minimize the impact of phishing attacks.great value to the overall security of an organization.

PAPER - 1

Intelligent web-phishing detection and protection scheme using integrated features of Images, frames, and text

The combined features of the text, graphics, and frames are used in this paper's robust adaptive neuro-fuzzy inference system (ANFIS)-based web-phishing detection and protection approach. The integrated elements of phishing websites' graphics, frames, and text were used in this work to propose an intelligent phishing detection and defense strategy. Based on the plans outlined in Aburrous et al. (2010) and Barracouta et al. (2014), an effective ANFIS algorithm was created, examined, and verified for phishing website identification and protection

(2015). The study employs a hybrid technique to choose text features, utilizing factors such as Page rank, Google Index, long URLs, domain entity relationships, and many others.

PAPER - 2

A Deep Learning Technique for Web Phishing Detection Combined URL Features and Visual Similarity

- ✓ Listed ways in which phishing detection is usually performed
    1. DOM (Document Object Model) tree analysis
    2. Visual feature based technique
    3. CSS(Cascading Style Sheets) based similarity analysis
    4. Website image comparison
    5. Visual perception method
    6. Hybrid Method
- ✓ Use of Convolutional Neural Networks(CNNs) to detect web phishing attacks based on URLs and screenshots of websites , as CNN is best in processing high dimensional data such as videos and images

- ✓ Division of snapshots of legitimate and suspected pages into blocks and matching using Earth Mover's algorithm , which gives high detection rate(99.6%)
- ✓ Normalized Compression Distance (NCD) to compute similarities based on distance between the image of a requested website and the image of a cached benign website , which gives a high true positive rate but is practically impossible with real time browsing
- ✓ Formulation of web phishing detection as a binary classification problem with the URL and images of websites as input leading to their classification as either legitimate websites or phished websites which can detect newly created phishing webpages based only on the URL and the screenshot of suspicious websites.

PAPER - 3

Intelligent cyber-phishing detection for online

- ✓ Methodology combines blacklist-based features, web content-based and heuristic-based approaches enabling a set of data (phishing websites, suspicious websites, spoofed web and legitimate websites) to be extracted from diverse sources.
- ✓ Based on evaluation of proposed methodology using ANFIS and MATLAB, using randomised and time-based evaluation, the method achieved 98.1% accuracy with 1.9% average testing error. Upon fine tuning, 99% accuracy was achieved with a 0.1% average error. Time taken to build the model was 0.01 Secs. The results demonstrate that the method can generalise well to new phishing attack
- ✓ J48 classifier shows a good performance that is 99.3% instances are correctly classified (TP), while 0.66% instances are incorrectly classified. Time taken to build the model is 0.01 secs
- ✓ NB classifier, on applying Randomised and Time-based evaluation methods, has high performance. 99.3% phishing instances correctly classified (TP) and 0.66% instances are incorrectly classified, while Time taken to build the model was 0.01 Secs. Recall maintained the accuracy, while F-Measure average score

slightly decreased by 0.03%. ROC curve average score decreased further by 0.5%, Precision score was the lowest with 98.7% accuracy. The overall accuracy (99.3%) exceeds 95%

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 behavioral response. As shown in the previous sections, most of the educational materials were also associated with a decrease in the *T N* rate, with an exception of only one educational material, namely *Anti-Phish Phil*. This shows that the addition of user training approaches is not *always* the right answer.

User education materials can complement software solutions. However it should also be noted that none of the existing studies empirically show enough evidence that user education can practically complement software solutions. This is due to the fact that all of the publicly available user education studies have evaluated educational materials independently from software solutions. This survey reviewed a number of anti-phishing software techniques. Some of the important aspects in measuring phishing solutions are:

• Detection accuracy with regards to zero-hour phishing attacks. This is due to the fact that phishing websites are mostly short-lived and detection at hour zero is critical.

• Low false positives. A system with high false positives might cause more harm than good. Moreover, end-users will get into the habit of ignoring security warnings if the classifier is often mistaken.

Generally, software detection solutions are:

• Blacklists.

• Rule-based heuristics.

• Visual similarity.

• Machine Learning-based classifiers.

The Machine Learning-based detection techniques achieved high classification accuracy for analyzing similar data parts to those of rule-based heuristic techniques.

REFERENCES:

[1] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," *IEEE Trans. Dependable Secur. Comput.*, vol. 3, no. 4, pp. 301–311, Oct. 2006.

[2] B. Krebs, "HBGary Federal hacked by Anonymous," http://krebsonsecurity.com/2011/02/hbgary-federal-hacked-by-anonymous/, 2011, accessed December 2011.

[3] B. Schneier, "Lockheed Martin hack linked to RSA's SecurID breach," http://www.schneier.com/blog/archives/2011/05/lockheed martin.html, 2011, accessed December 2011.

[4] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in *NDSS '10*, 2010.

[5] X. Dong, J. Clark, and J. Jacob, "Modelling user-phishing interaction," in *Human System Interactions, 2008 Conference on*, may 2008, pp. 627-632.

Andrew Jones After 25 years' service with the British Army's Intelligence Corps during which he was awarded an MBE, he became a manager and a researcher and analyst in the area of Information Warfare and computer crime at a defense research establishment. He developed and managed a well-equipped Computer Forensics Laboratory and took the lead on a large number of computer investigations and data recovery tasks. He holds a Ph.D. in the area of threats to information systems. He has written Fi five books on topics including Information Warfare, Risk management and Digital forensics and Cyber Crime and is currently writing a book on digital forensic procedures and practices.