ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

A Survey on Phishing Detection and Prevention Technique

Gaurav Kothari, Krishna Kumar, Rakhi Gulwane, Meghana Patil

B.E Student, Dept. of I.T, D.Y.P.I.E.T, Pimpri, Pune, Maharashtra, India.

ABSTRACT: Phishing has been the most easy and effective way to steal information from the user of internet. Phishing is a security attack on users of any website to obtain their trust and to get their personal and confidential information by presenting themselves a trustful source .They copy the appearances of website to obtain the private information. In this paper we discuss the existing phishing detection methods and algorithms form which one can be prevented from the attack of phishing. We discuss process of “Phishzoo”, in which the phishing website is detected by using profiles of trusted websites which is approx. 96% accurate .”Phishnet ”is the another method in which the author of the paper detect the phishing website by checking the URL of websites .Another paper using lexical features for detection of phishing URL is “Phishdef”. ”Phishstorm ” is next paper that we studied .The author of this paper proposed the new concept of intra-url relatedness used to evaluate the features extracted from words that compose a URL.

KEYWORDS: Profile,SIFT,online learning and offline learning

I. INTRODUCTION

Phishing is derived from 'fishing' and Fising refers to the act that the attacker allure users or visitor to visit a faked Sites by sending them faked e-mails and try to hacked personal information such as user name, password, and ID of national security etc. This information are used for further even identity theft attacks(Financial account). In these e-mails, they will make up some causes e.g. the password of your credit card and Debit card had been mis-entered for many times. to allure you visit their Web site to conform or modify your account details such as account number and password through the hyperlink provided in the e-mail. If you correct information as a input such as account number and password, the phishing attackers then successfully collect the information at the server side, and is able to perform their next step actions to hacked your information and able to withdraw money out from user account. There are various technologies that can be used for Phishing detection and prevention techinique. PhishZoo, PhishStrome, PhishNet are some of them this techinique mainly based on DNS detection and Content based technique.
Phishing Mostly used by phishers to steal user information and account to perform business crime in recent 20 years . by analysis it is prove that Within one to two years, the number of phishing attacks increased dramatically. The common characteristics in phishing hyperlink e-mails as listed below:
1) The actual link and the visual link are not the same;
2) The attackers use IP address with dotted decimal instead of DNS name;
3) Special tricks are used to encode the hyperlinks maliciously;
4) The phishing attackers often use fake DNS names that are same (but not identical) with the target or aimed website..

II. RELATED WORK

2.1 Phishnet[3]

Phishnet is new process of phishing detection.It contain two component which can help for phishing detection. URL prediction component:-In URL prediction component it generate new URLs (child) from known phishing URLs by employing various heuristics and test whether the new URLs generated are indeed malicious.Approximate URL matching component:-In approximate URL matching component perform an approximate match of a new URL with the existing blacklist.
Component 1: Heuristics for Generating New URLs

DOI: 10.15680/IJIRCCE.2016. 0401034

There are five heuristics for generating new URL

H1: Replacing TLDs

H2: IP address equivalence
H3: Directory structure similarity
H4: Query string substitution
H5: Brand name equivalence
Heuristics for Generating New URLs

H1: Replacing TLDs:-In heuristic 1 there are 3, 210 effective top-level domains (TLDs) and changes the effective TLD of the sorce URL with 3, 209 other effective TLDs.

H2: IP address equivalence:-In heuristic 2 phishing URLs having same IP addresses are grouped together into clusters and create new URLs by considering all combinations of hostnames and pathnames .

H3: Directory structure similarity:-In heuristic 3 URLs with similar directory structure are grouped together then build new URLs by exchanging the filenames among URLs belonging to the same group.

H4: Query string substitution:- In heuristic 4 it build new URLs by exchanging the query strings among URLs.

H5: Brand name equivalence:-At the last in heuristic 5 build new URLs by substituting brand names occurring in phishing URLs with other brand names.

Component 2: Approximate URL matching component
In 2nd component approximate URL matching there are 5 matching process are done given below .
M1: Matching IP Address:- In 1st matching it perform a direct match of the IP address of URL with the IP addresses of the blacklist entries and assign a normalizedscore based on the number of blacklist entries that map to a given IP address and if IP address IP_i is common to n_i URLs $\min\{n_i\}$ ($\max\{n_i\}$): the minimum (maximum) of all of phishing URLs which are stored in blacklisted entries of IP addresses .

M2: Matching Hostname:- In 2nd matching it perform hostname match with those in the blacklist and domains of phishing URLs Specifically registered for hosting phishing sites .

M3: Matching Directory Structure :-In 3rd matching it perform directory structure match with those in the blacklist.
1)Philosophy of this design
H3 (directory structure similarity)
H4 (query string substitution)
2)M4: Matching Brand Names:- In last matching it check for existence of brand names in pathname and query string of URLs

2.2 Phishstrom[1]

This paper propose an automated real-time URL phishingness rating system to protect users against phishing content: PhishStorm.We extract 12 features from a single URL which are input to machine learning algorithms to identify phishing URLs.
Phishingness score is computed for every single URL based on Random Forest classifier.We introduce the concept of intra-URL relatedness depicting the relation between a registered domain and the words that compose the rest of a URL.We leverage search engine query data to establish relatedness between words and show that this is more suited to Internet vocabulary than existing methods. We propose new features based on intra-URL relatedness and build a machine learning based approach relying on these for distinguishing between phishing and nonphishing URLs.

Type I: URL obfuscation with other domain: Thisis a real domain name, usually registered by thephisher, while the original domain being phished is part of the path, the query or the upper level domain.

Type II: URL obfuscation with keywords: Again the its is a real domain name, and the brand being phishedand related words are part of the path, the query or upper level domain.

Type III: Typosquatting domains or long domains: the URL is the domain being phished but misspelled, with letters or words missing or added, or the domain is pronounced the same way as the original but written differently. The targeted brand can also be com-bined with other words to create an unregistered domain.

Type IV: URL obfuscation with IP address: the URL'shostname is replaced by an IP address and the brand being phished is part of the path or the query.

Type V: Obfuscation with URL shortener: A URLshortening service is used to hide the name of the real host. Such URLs are not meaningful and are mainly used in phishing attacks targeting services that use this kind of short URL, like Twitter.

PhishStorm gives a generic solution for phishing URL de-tection relying on intra-URL relatedness computation. This technique only needs access to search engine query data to operate. Hence the application range of PhishStorm is wide. It can operate at different network level to prevent phishing. It can provide a personal protection for users while surfing on the Web if implemented locally as a browser add-on. PhishStorm provides phishingness score for URL and can act as a Web site reputation rating systems, displaying a Web site rating while using a search engine or typing a URL into a Web browser. Centralized phishing protection is an other option as for instance at the Web proxy level of a local company network, filtering HTTP packets sent from URLs identified as phishing.

However, as the main vector of phishing attacks is spoofed emails embedding phishing URLs, we implement PhishStorm as a centralized phishing email detection tool positioned in front of the email server. Nowadays, spam filtering is performed centrally in many organization and PhishStorm can be added to such process to increase detection performance. Fig. 2 depicts the implementation of PhishStorm and the four steps of the phishing email detection process. While incoming emails from the Internet reach PhishStorm , potential embedded URLs are extracted therefrom. The system then proceeds to features computation thanks to search engine query data and predicts a phishingness score using machine learning techniques . A detection threshold is applied to every predicted score, determining if the email must be forwarded to the email server and then to users with its phishingness score or dropped.We give in this section a detailed description of the implementation of the features computation process described in previous section. Phishstrom was important technique to prevent users from phishers.

2.3 Phishzoo[4]

In this part, we discuss that what is the approach of "phishzoo " paper .In this paper author explained their phishing detection approach. He start with an overview of the approach, followed by a explanation of site profiling and profile matching. Phishzoo can be worked on the online and offline phishing detection.
In this approach the phishing sites is detected by finding the content similarities between real site and the malicious site. Malicious sites tends to use sensitive sites. Phishzoo makes profile of sensitive websites and later use it for comparision between real and malicious site . There are advantages of this approach . First , the detection of phishing site depends on the current content of the website. It makes the process faster. Second , It works where URL-based machine learning approach fails . Third , as we know user provide there important sensitive information to few websites ,so to protect them phishzoo provides user-customized phishing protection to protect the information . It is also capable of detecting new attacks which were difficult for any other anti-phishing approach.

2.3.1 Profile:-

In a Profile phishzoo stores SSL certificate URL and content related to a site's appearance such as HTML files, extracted features of the logo. SIFT(Scale Invariant Features Tranform) algorithm is used to extract features from a site logo. This is used to transform a image in collection of local features vectors. We know that many sites uses logos which are scaled or translated version of some original logos. By using SIFT similarities between logos can be detected even in these cases.

2.3.2 Profile Matching:-

In this stage phishzoo checks if the present URL matches with the any whitelisted URLs. At first the hostname and HTML file is extracted .Some specific keywords are taken out and for them token are searched and for this TF-IDF technique is used.

2.3.3Image Matching using SIFT:-

SIFT is used by the computer application to recognize object in cluttered ,real world scence.Scale invariant features are needed in this case as because many phishers use the scaled translated images in their website which are hard to recognize by humans .Most current techniques fail to detect the picture in picture phishing attacks in which screenshot are used I phishing site of real site instead of html content .SIFT is used to overcome these problems .Before coming to SIFT, author found simpler image matching algorithms which is based on fuzzy hashing named as ImageMagick. This algorithms is faster than SIFT, but attacker can easly hack this technique. Author also found OCR algorithms, since many logos contain text. This worked very well in cases where the logo contains only text, but it failed when author studied more difficult logos such as the logos of EBay and Bank of Amer ica. Author determined that a more strong , vision-based approach was needed. SIFT image matching is a standard approach that is used in many object recognition and researches . Many researches used on variants of SIFT to improve matching speed in applications . These variants or a customized variant might prove good for anti-phishing research.

2.3.4 Running PhishZoo in Bulk:-

Authors analysis shows PhishZoo as a tool that will be used to protect end-users against phishing attacks. However, authors approach was more useful to intermediaries, like there are portals, browsers, ISPs, law enforce and security companies, who want to collect phishing sites for the purposes of blacklisting, takedown, or research. These intermediaries can run a PhishZoo that includes many more profiles from links in emails, webcrawling, or advertisements .This process may enable faster detection than the other sourcing techniques commonly relied upon.

2.3.5 Online and offline profile matching:-

Note that thousands of phishing attacks are happening everyday and phishing trends change quickly, however, according to phishtank within the time frame of our experiment the sites we chose to profile had more reported phishing attacks than other sites. Within any site, we made profile of the page that asks for confidential information, for example account number, password, PIN number, user ID. We also limited our analysis to sites that support SSL. In our dataset, 18% of the phishing sites had identical hash values. It is likely that some of these identical sites represent a single attack hosted across multiple domains (as in the Rock Phish attacks described by Moore and Clayton [31]), however, others represent distinct attacks that simply copy sites wholesale from the original page or other phishing attacks. As the numbers of duplicates we found were significantly lower than the 50% reported in that study, we suspect Phishtank has improved their filtering and decided to include these sites in our results. According to our manual analysis, 77.36% of the phishing sites in our dataset look similar to some real sites. 21.07% of the sites represent some real sites but the real site has no such page, for example an account confirmation page for Paypal but the real Paypal site has no such page, or a fake “claim your award” page for Bank of America. 1.57% of the phishing sites do not represent any real sites. These are free offer sites that ask for bank account numbers or other credentials.Phishzoo was important technique to prevent users.

2.4 PhishDef[2]

CLASSIFICATION ALGORITHM:-

Notation that we used: Denote features of an URL as a vector x and it is labeled as $y \in \{1,-1\}$, where 1 means URL is malicious and if it not the malicious then denoted by -1. Given that the new data x, and the goal is to predict the label y. The product between w and x is predicted by h(x),
$h(x) = \text{sign}(w \cdot x)$

Batched-based vs Online.

A Batched-based algorithm train its model which is based on batched labeled data. New data is predicted by this training model. It is because training the model of batch based algorithm required the batch of data. Online algorithms require significantly less memory than batch-based algorithms.

A. Batch Learning

Support Vector Machine (SVM): It is widely known for obtaining accurate classification of high- dimensional data. They also perform well in the area of classifying malicious URLs . In this paper, we focus on the performance of batch-based SVMs.

B. Online Learning

The online algorithms operate in rounds. In round t, receives xt and it predicts xt's label as ^yt using the online learning model; then received the true label yt, and also updates its model based on (xt,yt).

1) Online Perceptron (OP): If errors comes OP updates w continuously. Or can say that, w is updated only if the predicted label, ˆ yt = sign(wt · xt), disagreed with the true label, yt, of xt. The update is as follows:
wt+1 ← wt + yt xt .
OP has some drawback: the rate of updation is fixed and does not count the magnitude of classification error, which leads to poor accuracy.
2) Confidence Weighted (CW): CW which is a binary classification algorithm, introduced by Author.CW acquired the notation of confidence in the term of a feature. With this confidence notion, CW point out the drawback of Online Perceptron through two mechanisms: (i) The weights of the more confident features of the OP is updated by CW less aggressively; and (ii) CW does not change the weights too much but it is just enough to correct for the mistake.
3) Adaptive Regularization of Weights (AROW): This is the last algorithm in this category that we examined is AROW algorithm by Author. AROW can be considered as the next level of CW so that the classifier is more build in the presence of label noise. Let us an example, if 'parliament.in' is wrongly labeled as the malicious (by an adversary) and supplying to CW, then the CW will make some changes to all the features which are contained in the URL , if it sees this URL again, it will be make the flag to this URL as malicious.

CW, therefore, increase the values of the feature "top level domain is .in". AROW avoids this strong behave by softening the formulation of CW. Formally, Author develop the motive to CW as regularizers. The update rule is as follows:

(μt+1,Σt+1)= arg min DKL(N(μ,Σ)||N(μt,Σt)) + λ1lh^2(yt,μ · xt) + λ2xt^T Σxt ,
Where lh^2(yt,μ·xt) = (max{0,1−yy(μ·xt)})^2

is the squared- hinge loss suffered using μ to predict the label for xt when its true label is yt, and λ1 and λ2 are parameters.It is similar to CW, the update running time is linear in the number of non-zero features in xt. The requirement of memory is constant in terms of the input data. As we know that In phishing context there are various methods or the algorithms used. But it is the first time AROW is used.

III .CONCLUSION AND FUTURE WORK

Serious network security problem is Phishing , causing billions of dollars financial lose to both consumers and e-commerce companies. Phishing has made lot of distrusted and less attractive to e-commerce, consumers and perhaps more fundamentally, the various techniques is available for detecting and preventing the phishing sites and e-mail. In this paper, we have studied the characteristics of the hyperlinks that were embedded in phishing e-mails. We come with many anti-phishing techniques. Since Phishing Guard is characteristic based, it can not only detect known attacks, but also is effective to the unknown ones. We show many technique have there advantage and disadvantages. PhishZoo have 94% detection probability. But the algorithm used in PhishZoo is heavy weighted. In this paper we showed that many technique along with there false positive and false negative. PhishNet contain two component for Phishing detection and prevention such as URL prediction component and URL matching component. Phishzoo can be worked on the online and offline phishing detection.

In PhishStrom propose an automated real-time URL phishingness rating system to protect users against phishing content: PhishStorm. We extract 12 features from a single URL which are input to machine learning algorithms to identify phishing URLs. In PhishDef it uses classification algorithm for phishing site detection and prevention. In our future scope we use Link Guard algorithm which is light-weighted and can detect up to96% unknown phishing attacks in real-time. In this paper we believe that Link Guard is not only useful for detecting phishing attacks, but also can prevent shield users from malicious or unsolicited links in Web pages and Instant messages.
Link Guard algorithm is most effective and light-weighted for detecting Phishing sites and preventing from phishing sites.

REFERENCES

[1]Samuel Marchal, Jérôme François, Radu State, and Thomas Engel "PhishStorm: Detecting Phishing With Streaming Analytics"
[2] Michalis Faloutsos University of California, Riverside michalis@cs.ucr.edu, Anh Le, Athina Markopoulou University of California, Irvine {anh.le, athina}@uci.edu, "PhishDef: URL Names Say It All"
[3] Pawan Prakash, Manish Kumar, Ramana Rao Kompella, Minaxi Gupta Purdue University, Indiana University, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks".
[4] Sadia Afroz Department of Computer Science Drexel University Philadelphia, PA 19104 Email: sa499@drexel.edu, Rachel Greenstadt Department of Computer Science Drexel University Philadelphia, PA 19104 Email: greenie@cs.drexel.edu "PhishZoo: Detecting Phishing Websites By Looking at Them".

BIOGRAPHY

Gaurav Kothari,Rakhi Gulwane,Krishna Kumar and Meghana Patil perceiving degree in Pune University.They are working on project named "Phishing detection and prevention technique –Link Guard Algorithm".