ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

An Innovative Framework for the Detection and Prediction of Phishing Websites

Divya James

Assistant Professor, Dept of Information Technology, Rajagiri School of Engineering and Technology, Kochi, Kerala, India

ABSTRACT: With the advent of internet, various online attacks has been increased and among them the most popular attack is phishing.Phishing is an attempt by an individual or a group to get personal confidential information such as passwords, credit card information from unsuspecting victims for identity theft, financial gain and other fraudulent activities. In recent years phishing is a technique used for cyber crimes. Spoofing is a new type of cyber crime in this globalisation era. Spoofing refers tricking computer systems or computer users by hiding one‟s identity or faking the identity of another user on the Internet.E-mail spoofing means sending messages from a bogus e-mail address or faking the e-mail ID of another user. This paper employs back propagation network for identifying malicious URL‟s in a network .It has been observed that the method predicts the phishing website more accurately when compared to any other learning algorithms.

KEYWORDS: Phishing ,Classification ,Artificial Neural Network, Machine learning

I. INTRODUCTION

Online transactions are nowadays become very common and there are various attacks present behind this. In these types of various attacks, phishing is identified as a major security threat and new innovative ideas are arising with this in each second so preventive mechanisms should also be so effective. Thus the security in these cases be very high and should not be easily tractable with implementation easiness. Today, most applications are only as secure as their underlying system. Since the design and technology of middleware has improved steadily, their detection is a difficult problem. As a result, it is nearly impossible to be sure whether a computer that is connected to the internet can be considered trustworthy and secure or not. Phishing scams are also becoming a problem for online banking and e-commerce users. The question is how to handle applications that require a high level of security. Phishing[1] attacks rely upon a mix of technical deceit and social engineering practices. In the majority of cases the phisher must persuade the victim to intentionally perform a series of actions that will provide access to confidential information. Communication channels such as email, web-pages, IRC and instant messaging services are popular. In all cases the phisher must impersonate a trusted source (e.g. the helpdesk of their bank, automated support response from their favourite online retailer, etc.) for the victim to believe. To date, the most successful phishing attacks have been initiated by email – where the phisher impersonates the sending authority (e.g. spoofing the source email address and embedding appropriate corporate logos). For example, the victim receives an email supposedly from support@mybank.com (address is spoofed) with the subject line 'security update‟, requesting them to follow the URL www.mybank-validate.info (a domain name that belongs to the attacker – not the bank) and provide their banking PIN number.

Phishing performed through five aspects and are discussed below:

1. Try to obtain the email address of the target by several ways.
2. Attackers create the websites that are very similar to the legitimate websites.
3. Then they send the link to target email address for accessing created fake website.
4. Normal users were clicking the link and get trapped.
5. Fake website catches the user credentials and use their secret information to steal money and identity from the victims

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

personal accounts [2].

One of the Phishing technique used by the attackers includes: Phishing through Compromised Web servers –If a server is compromised then the attacker used that server as a tool for attacking other systems. Also if a server is compromised then a root kit or password protected backdoor can be installed by the attacker. From that onwards they can use it as a legitimate user credentials.

II. LITERATURE SURVEY

In 2012, Gaurav, Madhuresh Mishra , Anurag Jain proposed an Anti-Phishing Technique Using Pattern Matrix which keeps users away from phished websites. They proposed a prevention based technique by which each website require user credentials for accessing it instead of using the hyperlinks. The users can access the website from anywhere by setting authentication using code generation and hashing [3]. Juan Chen,Chuanxiong Guo presented Link Guard based online detection and prevention of phishing attacks on WindowsXp.They designed Link Guard algorithm not only for detecting phishing but also it resist users to click on malicious and unsolicited links.The system detects the phishing up to 96% [4]. Engin Kirda and Christopher Kruegel Technical University of Vienna, proposed an anti phishing techniques using AntiPhish algorithm**.** This technique tracks the sensitive information of a user and generates warnings whenever the user attempts to give away details to a web site that is considered untrusted. It is used to check the trustworthiness of a web site. The system mainly focused on web based attacks.[5] A paper titled as “Anti-phishing Based on Automated Individual White-List” by Ye Cao, Weili Han, Yueran Le introduced a novel anti-phishing approach named Automated Individual White-List (AIWL). Automated Individual White-List (AIWL) maintains an information list about the familiar users of the websites. If an attacker tried to access the websites then it checks the list, if he is not in the list then the system alert the websites about the attacks. AIWL is effective in detecting phishing and pharming attacks with low false positive. For maintaining the white list the system uses “The Naive Bayesian” classifier .The system provides more accurate alerts to the users.[6]

III.PROPOSED SYSTEM

A.Phishing Website Features There are several features that distinguish phishing websites from legitimate ones. In our study, we used 18 features descried briefly hereunder:

1 .IP address: Using IP address in the hostname part of the URL address means user can almost be sure someone is trying to steal his personal information.

2. Long URL: Phishers resort to hide the suspicious part of the URL, which may redirect the information submitted by the users or redirect the uploaded page to a suspicious domain.

3. URLs having “@” symbol: The “@” symbol leads the browser to ignore everything prior it and redirects the user to the link typed after it.

4. Prefix and Suffix in URLs: Phishers deceive users by reshaping the URL to look like legitimate ones. A technique used to do so is by adding prefix or suffix to the legitimate URL so users might not notice any difference.

5. Sub-domain(s) in URL: Another technique used by the phishers to deceive the users is by adding sub-domain(s) to the URL thus the users may believe that they are dealing with a credited website.

6.Misuse of HTTPs protocol: The existence of the HTTPs protocol every time sensitive information is being transferred reveals that the user certainly connected with an honest website. However, phishers may use a fake HTTPs protocol so that users might be deceived. In [7] a recommendation to check whether the HTTPs protocol is offered by a trusted issuer such as “GeoTrust, GoDaddy”.

7.Request URL: A webpage usually consists of a text and some objects such as images and videos. Typically, these objects are loaded to the webpage from the same domain where the webpage exists. If the objects are loaded from a domain different from the domain typed in the URL address then the webpage is potentially suspicious.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

8.URL of Anchor: Similar to "Request URL" but for this feature the links within the webpage might refer to a domain different from the domain typed on the URL address bar. This feature is treated exactly as "Request URL".
9.Server Form Handler "SFH": Once the user submits his information, that information will be transferred to a server to be processed. Normally, the information is processed from the same domain where the webpage is being loaded.Phishers resort to make the server form handler either empty or the submitted information are transferred to different domains.
10.Abnormal URL: If the website identity does not match its record shown in the WHOIS database (http://who.is/) the website is classified as "Phishy". This feature is a binary feature.
11. Redirect Page: This feature is commonly used by phishers by hiding the real link which asking users to submit their information to a suspicious website.
12. Using Pop-up Window: It is unusual to find a legitimate website that asks users to submit their credentials through a popup window.
13. Hiding the Suspicious Links: Phishers resort to hide the suspicious link by showing a fake link on the status bar of the browser or by hiding the status bar itself.
14. DNS Record: If the DNS record is empty or not found the website is classified as "Phishy", otherwise it is classified as "Legitimate".
15. Website Traffic: Legitimate websites are of high traffic since they are visited regularly. Phishing websites often a short life thus their web traffic is either not exists ranked is below the limit that gives it the legitimate status.
16. Age of Domain: the website is considered "Legitimate" if the domain aged more than 2 years [7].
17. Disabling Right Click: Phishers use JavaScript to disable the right click function, so that users cannot view and save the source code.
18. Port number: We examine if there is a port number in the URL and check if the port belongs to the list of well-known HTTP ports such as 80, 8080, 21, 443, 70, and 1080. If the port number does not belong to the list, we flag it as a possibly phishing URL.

B. Artificial Neural Network

An Artificial Neural Network (ANN) is an information processing model that is stimulated by how biological nervous systems process information. The key element of this model is the unique structure of the information-processing scheme. NN consist of a large number of highly interconnected processing elements "neurones", working in harmony to solve problems. ANNs, like human, learn by example. NNs, with their amazing ability to derive meaningful data from complex dataset, can be used to mine patterns that are too difficult to be noticed by humans [8]. A trained NN can be thought of as an "expert" in the domain it has been applied and can be used to predict class of new cases. Other advantages include [8]:
• Nonlinearity: NN is very effective technique in modeling classification problems where the output values are not directly related to its input.
• Adaptive: Neural network has the ability to adjust the weights based on the changes of its surrounding environments.
• Generalisation: NN is able to find the suitable output for the inputs that does not exist in the training data.
• Fault-tolerance: NN performance is not significantly affected under difficult circumstances such as losing connection between some neurons, noisy or missing data.
• Identical designing steps: The same principles, scheme and methodological steps are employed in designing ANN in all domains.

C Back Propagation Algorithm

The back propagation algorithm [9] is the most frequently used training method for ANNs.Backpropagation is usually implemented along with feed forward NNs that have no feedback. The main idea in feedforward NNs is to propagate the error through the hidden layers to update the weights of NN. The back-propagation algorithm is described as the following pseudo code:
Initialize the weights vector S = the training set fed to the network
Repeat For each "input-output" pair denoted by P in S

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

In = input pattern in P Out = desired output Compute network output (netout)
network error = Out – netout end For Find weight change for weights connecting hidden to output Find weight change for weights connecting input to hidden
Update weights Until reaching (a satisfactory network error value OR maximum iteration)

D Proposed URL Analyser

1. Lexical Features (F1)
Lexical features analyses the format of the URL. It includes the length of the host name, length of the URL, the number of dots, presence of suspicious characters such @ symbol, hexadecimal characters and other special binary characters such as („.“, „=“, „$“, „^“ and etc.) either in the host or path name. IP addresses and hexadecimal characters are used to hide the actual URLs. Consider the URL http://www.citibank.com@phishingsite.com” will enter into “phishingsite.com” and discards “www.citibank.com”. These kinds of techniques use the actual phishing website to disguise and pose as legitimate sites.

2 .Host Based Features
Host based features identify the location,owner and how malicious sites are hosted and managed. Some of the features are as follows:

2.1.Age of Domain
Age of the domain is used to identify when malicious websites are hosted such that they have less age or relatively new to obtain the user credentials. They will be recently registered sending more mails and some domains may not be available even at the time of checking. It obtains the data in the number of months and some may be in years more recently. The WHOIS lookups on the WHOIS server is used to retrieve the domain registration date, and if the domain registration entry is not found on the WHOIS server, this feature will simply return -1, deeming it suspicious.

2.2 Page rank(F3)
Page rank provides the rank for the webpage and proves higher the page rank, the more important is the page. Obviously phishing web pages have less age of domain and short lived. Hence they obtain a very low page rank or page rank does not exist. Page rank is a link analysis algorithm first used by Google, in which each document on the web is assigned a numerical weight from 0 to 10, with 0 indicating least popular and 10 meaning most popular. A score value of 1 is assigned when the page rank value for a particular webpage is not available.
3 Number of Sensitive Words in URL

3.1 Individual occurrences(F4)and Co-Occurencesof suspicious phishing keywords (F5)
Abu-Nimeh et al (2007) used the “bag-of-words” approach with a list of 43 most frequent words as features in a machine learning approach. Garera et al (2007) used a set of eight sensitive words such as secure, account, update, login, sign-in, banking, confirm and Verify that frequently appear in phishing URLs. The system is trained with 1000 phishing emails to give weights to the suspicious words found in the phishing emails.The count of most occurring words includes Secure, Account, Update, Login, Verify ,Signin, Banking,Notify, Click, Inconvenient, password etcand their Co-Occurences in the phishing mail. The classifier has a training dataset of malicious phishing URLs and legitimate URLs. The probability occurrence of each feature in the dataset are calculated and their respective scores are obtained (i.e) Count up occurrence of features in the dataset and calculate the cumulative score. If Cumulative score > Threshold, consider as phishing URL else legitimate URL.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

V. RESULTS AND CONCLUSION

Hackers bypass anti-spam filtering techniques by embedding malicious URL in the content of the messages. Hence the above method with the help of minimized phishing feature set identifies the malicious URL in the emails. The datasets are obtained from two sourcesviz DMOZ Open Directory Project and Phishtank(2012). Phishtank is a source of blacklisted phishing URLs which admits user inputs and they are also verified by users. The false positive rate refers to the number of legitimate emails classified as phishing emails,and false negative rate refers to the number of phishing emails classified as legitimate. The Table 1 shows that out of 1000 Phishing mails with malicious URLs, the above results were obtained for identifying various lexical and host based features. Table 1

Technique	Number of Features(n)	TPR (%)	FPR(%)
Cantina(Existing) with n1 features	n1(20)	89	1
Cantina +(Existing) with n2 features	n2 (27)	92.54	0.407
Proposed method with m features	m (14)	94	0.5

REFERENCES

[1]. Ollmann G., The Phishing Guide Understanding & Preventing Phishing Attacks, NGS Software Insight Security Research.
[2] http://www.w3.org/TR/xhtml1
[3]. Gaurav, Madhuresh Mishra, Anurag Jain / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 3, May-Jun 2012, pp.1825-1828
[4]. Online Detection and Prevention of Phishing Attacks (Invited Paper) Juan Chen, Institute of Communications Engineering Nanjing 210007, P.R. China
[5]. Protecting Users Against Phishing Attacks with AntiPhish ,Engin Kirda and Christopher Kruegel Technical University of Vienna.
[6] Y. Zhang, J. Hong and L. Cranor. CANTINA: A Content-Based Approach to Detecting Phishing Web sites.Proceeding of International World Wide Web Conference (WWW 2007), Banff, Alberta, Canada, May 2007: 639-648.
7] R. M. Mohammad, F. Thabtah and L. McCluskey, "An Assessment of Features Related to Phishing Websites using an Automated Technique," in *The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012)*, London, 2012.
[8]. I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," ACM, New York, NY, USA, March 2002
[9]. B. Widrow, M. and A. Lehr, "30 years of adaptive neural networks," *IEEE press*, vol. 78, no. 6, pp. 1415-1442, 1990.