Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 4 (2013) 125 – 131

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Intelligent Systems and Control

Detecting Malicious URLs in E-Mail – An Implementation

Dhanalakshmi Ranganayakulu[a], Chellappan C[b]*

[a]*Adhiparasakthi Engineering College, Melmaruvathur 603319, INDIA*
[b]*Anna University, Chennai 600 025, INDIA.*

Abstract

The World Wide Web has become the most essential criterion for information communication and knowledge dissemination. It helps to transact information timely, rapidly and easily. Identity theft and identity fraud are referred as two sides of cybercrime in which hackers and malicious users obtain the personal data of existing legitimate users to attempt fraud or deception motivation for financial gain. E-Mails are used as phishing tools in which legitimate looking emails are sent making the genuine users identity with legitimate content with malicious URLs. It helps to steal consumers' personal data such as user names, account numbers, passwords and other financial account credentials. Spam E-Mails emerges or transforms as Phishing mails. Spoofed Mails plays a vital role in which the hackers pretends to be a legitimate sender posing to be from a legitimate organization which divulges the user to give his personal credentials. The content may escape from Content based filters or the email may be without any body of the message except malicious URL in it. This paper identifies malicious URLs in email through reduced feature set method.

Keywords: Age of Domain;Host based features;Lexical features;Malicious URLs;Page rank; Phishing.

1. Introduction

The Web serves as better medium for a large number of malicious activities such as Spam attacks, Phishing attacks ,DDos attacks and etc. motivated under financial aspects. These attacks attract the common

* *E-mail address:* dhanalakshmisai@gmail.com.

doi:10.1016/j.aasri.2013.10.020

users to click links attached in legitimate looking or spam emails and make them to visit the malicious sites. It initiates them to click, urges them to give their personal information. Phishing attacks are referred as Lure, Hook and Catch (Jacobsson and Myers 2007). Spoofed E-Mails poses to be from legitimate company seeking sensitive information. These email addresses are called the 'Lure'. E-mails with malicious URLs may have legitimate content in the body of the mails which are unable to be detected by content based spam filters. The URLs lead to the actual Phishing sites which are clones of legitimate websites and lure the users into entering sensitive information. The actual phishing websites are the 'Hook' which obtains the private information from the user. The malicious user poses various critical conditions such as account suspension, failed transaction and forcing user to upgrade the newly installed security feature. The links in the email leads to fake phishing site referred as 'Catch'. The legitimacy of the website may not be displayed by the browser which outlooks the phishing websites as legitimate. In some cases, the user also overrides the browsers decision.

2. Existing solutions

Blacklists may be in the form of IP addresses or websites used by email filters and block the users through an available list of IP addresses or websites. PhishNet (Pawan et al 2010) enhances existing blacklists by discovering related malicious URLs. One major problem with blacklists is that they fail to identify phishing URLs in the early hours of a phishing attack because their update process is insufficiently fast. Phishing campaigns have an average life of less than two hours (Sheng et al 2009) and by the time a phishing website is positively identified and blacklisted, it would have almost hacked. Various features are extracted from URLs which includes suspicious characters, number of dots in the URL, hexadecimal characters, IP addresses and length of the URL. Colin Whittaker et al (2010) discussed a scalable machine learning algorithm to automatically classify phishing pages by training the classifier on noisy dataset. The false positive rate is below 1% and the classifier is based on Google's phishing blacklist URLs. Justin Ma et al(2009) discuss a method to detect malicious websites by analyzing lexical and host based features based on passive aggressive algorithm. The improvement can be obtained by analyzing the features of page content and pagerank. Zhang et al (2007) proposed a content-based method using a linear classifier and achieved 89% TP(True positive) and 1% FP(False positive).The test case was demonstrated for 100 phishing URLs and 100 legitimate URLs. CANTINA+ (2010) classifies phishing URLs and the feature set is more exhaustive and obtained a classification accuracy of 92.3%. There exist various related researches and case studies conducted on analyzing the feature set required to reduce the exhaustiveness and time consumption. Maher Abburrous et al (2010) attempted a survey to identify the required features which helps to improve the accuracy and the precision of detecting malicious URLs. Various sources of phishing attacks are obtained from APWG's archive(2011) and Phishtank archive(2012). The features are listed in Table 5.1.

Table 1 High impact features on URL phishing instances [Maher Aburrous et al (2010)]

S.No.	Phishing Features	No. of appearances	Appearance %
1.	Using the IP address	14	46.66✓
2.	Abnormal request URL	30	100✓
3.	Abnormal URL of anchor	7	23.33✓
4.	Abnormal DNS record	2	06.66
5.	Abnormal URL	5	16.66
6.	Using SSL certificate	17	56.66
7.	Certification authority	4	13.33
8.	Abnormal cookie	2	06.66
9.	Distinguished Names Certificate (DN)	4	13.33
10.	Redirect pages	3	10.00
11.	Straddling attack	2	06.66
12.	Pharming attack	4	13.33
13.	Using on MouseOver to hide the link	6	20.00
14.	Server Form Handler (SFH)	2	06.66
15.	Spelling errors	24	80.00✓
16.	Copying website	5	16.66
17.	Using forms with ‘‘Submit’’ button	6	20.00✓
18.	Using Pop-Ups windows	8	26.66✓
19.	Disabling right click	2	06.66
20.	Long URL address	22	73.33✓
21.	Replacing similar characters for URL	16	53.33✓
22.	Adding prefix or suffix	9	30.00✓
23.	Using the @ symbol to confuse	6	20.00✓
24.	Using hexadecimal character codes	8	26.66✓
25.	Much emphasis on security and response	5	16.66
26.	Buying time to access accounts	3	10.00

The above selected features (✓) display high impact in various studies as mentioned in the literature and hence the feature set comprises features whose impact is greater than 20%.This involved the host based features, lexical features,page rank and suspicious keywords in the mail for better performance.

3. URL analyzer

Phishing URLs can be analyzed based on the lexical features and host based features of the URL. The lexical feature analyses the format of the URL. URLs contain the host name and the path. For example, consider ‘www.annauniv.edu/emmrc/emmrc.html’, the host name is www.annauniv.edu and emmrc/emmrc.html is the path. The proposed methodology analyses host based features such as Pagerank and age of domain, various lexical based features such as URL encoding, presence of suspicious characters, hexadecimal character or malicious IP addresses to hide them and analyses the word probabilities to find whether the email contains any suspicious links to avoid end users falling by phishing attacks as illustrated in Fig 1. Itis useful as illegitimate users spoof their identities, pass authentication tests and during content analysis also it may get escaped by avoiding spam keywords. Some emails may not contain any message in the body except some malicious links in it urging the users to click them leading to fraudulent websites.

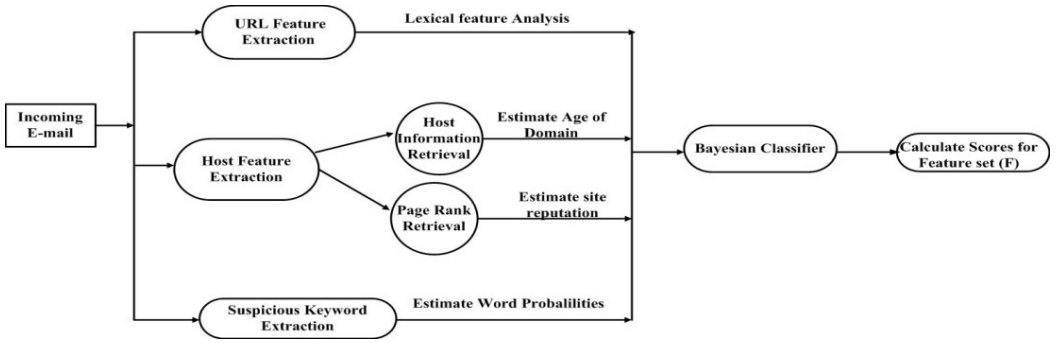

Fig 1 URL feature extraction

3.1 Lexical Features (F_1)

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. For example consider the URL http://www.bankingcompany.com/online /transaction/website/phishing.html" which is shortened using the IP address http://132.115.201.115 which looks like legitimate and not suspicious. The URL can also be represented using hexadecimal base values with a '%' symbol. It may represent any special characters Spoofguard (Neil Chou et al 2004) identified the '@' and '-' symbol most prominent in phishing URLs. A @ symbol in a URL will enable the URL at the left to discard which is legitimate URL and right to enter into the phishing site. Consider the URL http://www.citibank.com@phishingsite.com" will enter into "phishing site.com" and discards "www.citibank.com". These kinds of techniques use the actual phishing website to disguise and pose as legitimate sites.

3.2 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

3.2.1 Age of domain (F_2)

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.

3.2.2 Page rank(F_3)

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. After examining the dataset, 1000 phishing mails and 1000 legitimate mails the percentage of emails matching the Lexical and Host based features are listed in Table 2.

Table 2Characterizing lexical and host based features matching with Phishing Mails

Feature	Legitimate	Phishing
Has IP Address	0%	0.04%
Has "Hexadecimal" Character	0%	0.01%
Has suspicious character '@' symbol	0%	0.01%
More No. of Dots	0.01%	0.06%
Suspicious Age of Domain	35%	75%
Page rank< 3 feature	1.2%	88%

3.3 Number of Sensitive Words in URL

3.3.1 Individual occurrences(F_4)and Co-Occurencesof suspicious phishing keywords (F_5)

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 e-mails. The count of most occurring words includes Secure, Account, Update, Login, Verify ,Signin, Banking, Notify, Click, Inconvenient, password etcand their Co-Occurencesin the phishing mail.

*3.4Approach -Bayes Classifier*Bayes classifier is adapted in spam filters such that individual features of URLs are distributed independently of the values of other features. Bayes theorem is used to calculate the probability of hypothesis for the event B, provided with the training data A,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1.1}$$

It is often easier to calculate the probabilities, $P(A|B)$, P(A), P(B) for the probability$P(B|A)$that is required. Extrapolating Baye's rule, assume that legitimate and phishing websites occur equal in number and hence with equal probability, then the posterior probability that the feature vector X belongs to a malicious URL is such that

$$P(B = 1|A) = \frac{P(A|B=1)}{P(B|A=1)+P(B|A=0)} \tag{1.2}$$

$$P(B|A) = \frac{P(A|B)}{P(A|B)+P(A|B')} \tag{1.3}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B')P(B')} \tag{1.4}$$

where, P(A) = Probability of feature F in phishing and legitimate dataset.

$P(B') =$ Legitimate dataset.

$P(B) =$ Phishing dataset.

P(B(Phishing)) = P(B'(Legitimate)) = 0.5

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 as illustrated in Fig2 .

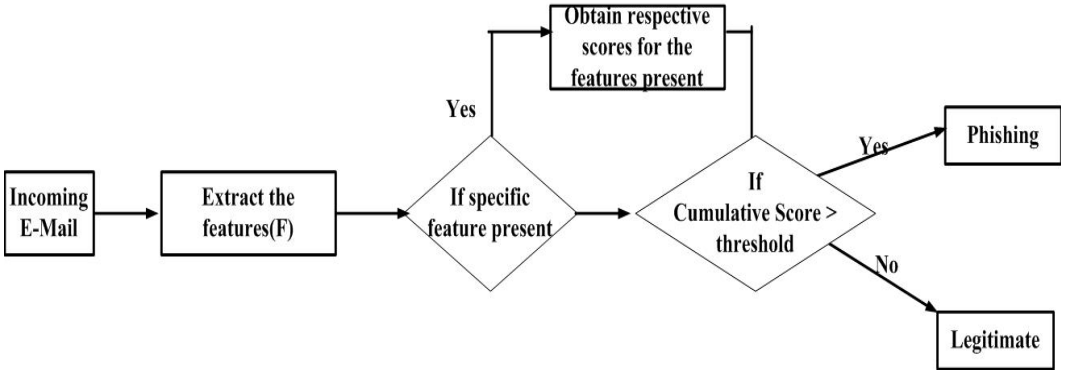

Fig 2 Phishing URL classifications

a) How many times does feature F(F_1,F_2,F_3,F_4,F_5,F_6) appear in phishing dataset?

b) How many times does feature F(F_1,F_2,F_3,F_4,F_5,F_6) appear in legitimate dataset?

Let F_1 = Lexical features ,F_2 = Age of the domain factor of URLs

F_3 = Occurrence of Pagerank< 3 in phishing and legitimate dataset

F_4 = Individual Occurrence of suspicious keywords

F_5= Co–Occurrences of suspicious keywords, F_6 = Login Form detection

4.Conclusion and Results

Hackers bypass anti-spam filtering techniques by embedding malicious URL in the content of the messages. Hence the URL analyzer 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. An E-Mail server has been configured with hMail namedas SSE Mail Server for the testing purposes. 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 3 shows that out of 1000 Phishing mails with malicious URLs, the above results were obtained for identifying various lexical and host based features.

Table 3 Performance analysis with the existing systems

Technique	Number of features	TPR (%)	FPR (%)	Time Complexity

	(n)			
Cantina (Existing) (with n1 features)	n1(20)	89	1	O(n1)
Cantina+ (Existing)(with n2 features)	n2(27)	92.54	0.407	O(n2) (n1<n2)
URL Classifier (Proposed)(with m features)	m(14)	92.8	0.4	O(m)(m<n2)

References

1. Colin Whittaker, Brian Ryner and MarriaNazif, "Large-Scale Automatic Classification of Phishing Pages", In proceedings of NDSS, 2010.
2. Fette, I., Sadeh, N. and Tomasic, A. "Learning to Detect Phishing Emails' In WWW", Proceedings of the 16th International conference on World Wide Web, pp. 649-656, 2007.
3. Garera, S., Provos, N., Rubin, A.D. and Chew, M. "A Framework for Detection and Measurement of Phishing Attacks" In Proceedings of the 2007 ACM workshop on Recurring malcode, pp. 1-8, 2007.
4. Justin Ma, Lawrence K. Saul, Stefan Savage and Geoffrey M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining pp.1245-1254, 2009.
5. Jacobsson, M. and Myers, S. "Phishing and Countermeasures - Understand the Increasing Problem of Electronic Identity Theft", New Jersey: Wiley, 2007.
6. Justin Ma, Lawrence Saul, K., Stefan Savage and Geoffrey Voelker, M. "Identifying Suspicious URLs: An Application of Large-Scale Online Learning", In ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 681-688, 2009.
7. Maher Aburrous, Hossain, M.A., KeshavDahal and FadiThabtah, "Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies", Cognitive Computing, DOI 10.1007/s12559-010-9042-7, Vol. 2, pp. 242-253, 2010.
8. Neil Chou, Robert Ledesma, Yuka Teraguchi, Dan Boneh and John Mitchell, "Client-side defense against web-based identity theft", In 11th Annual Network and Distributed System Security Symposium (NDSS '04),San Diego, 2004.
9. PawanPrakash, Manish Kumar, RamanaRaoKompella and Minaxi Gupta, 'PhishNet:Predictive Blacklisting to Detect Phishing Attacks', Proceedings of the IEEE Infocom, pp.1-5, 2010.
10. Sheng, S.,Wardman, B.,Warner, G., Cranor, L., Hong, J. and Zhang, C. "An empirical analysis of phishing blacklists", In Proceedings of the CEAS'09, 2009.
11. Xiang, G., Hong, J., Rose, C. P. and Cranor, L. "CANTINA+: A feature-rich machine learning framework for detecting phishing Web sites". ACM Trans. Inf. Syst. Secur. Vol.14, No.2, pp.1-21, 2011.
12. Zhang, Y., Hong, J. and Cranor, L. Cantina: A Content-Based Approach to Detecting Phishing Web Sites. In Proceedings of the 16th international conference on World Wide Web, pp.639-648, 2007.