March - April 2020
ISSN: 0193-4120 Page No. 8822 - 8830

An Efficient Approach for URL Phishing Attack Detection using Random Forest Algorithm

[1]Dr. S. Kanaga Suba Raja, [2]K.Valarmathi, [3]Naga Suryanarayana. D, [4]Niranjan. S,[5]Kavin Sathya. B

[1,2,3,4,5] Department of Information Technology, Easwari Engineering College
[1]sathya0598@gmail.com, [2]sniranjansp@gmail.com, [3]naga.surya1909@gmail.com, [4]skanagasubaraja@gmail.com, [5]Valar.me964@gmail.com

Article Info
Volume 83
Page Number: 8822 - 8830
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 09 April 2020*

Abstract

Phishing is a cyber-attack where attackers aim to steal user's personal information, login credentials and passwords, bank account details, location, etc., from naive internet users. Phishing attacks are the leading cause to information theft and other financial information theft. Both companies and daily users of the internet are affected by this malicious practice which illegally steals user information. We propose a machine learning approach to detect online phishing attacks using Uniform Resource Locator (URL) features. In this system, we have considered about 12 URL attributes to determine whether a website is benign, spam or malicious. The system is trained using about 4000 phishing and legitimate URLs using SVM and Random forest data classifiers. Our system is able to detect the nature of the website of up to 90% accuracy using SVM data classifier.

Keywords; *Phishing, URL, SVM, Random forest, Machine learning.*

I. INTRODUCTION

Phishing attacks are a major concern among individuals and major corporate companies. Malicious website deceives the user into giving their personal information willingly. Attackers usually copy the style and content of the original brand websites and make a copy of the original website where users are tricked into entering their personal information and other details which leads to loss for both the company's reputation and the compromise of individual's sensitive information. Phishing attacks usually take place via emails and copies of shopping websites and bank websites. Clicking on a malicious email automatically downloads malware and tries to obtain login credentials, while phishing websites trick the user to enter credit card details posing as a legitimate website. Attackers first analyse the legitimate websites and try to recognise the distinguishing features about the original website. The attackers then try to create their own phishing websites that almost resembles the original website. Although some features cannot be replicated, some obvious signs can be left out for users and developers to identify and protect themselves from phishing websites. The main objective of this paper is to identify such malicious websites and protect users from information theft. Attackers illegally obtain this information by tricking the users into entering their information on their malicious websites. Phishing websites are one of the main reasons for the increase in criminal web activities and deny web services. We propose a system where the implementation works on the name of the URL rather than the contents of the website which are java scripts and html tags. This is also known as natural language processing where mining is done on the text where the text here is the name of the websites we enter. We can also use factors like java scripts and html tags but using the URL's name alone will decrease the processing time of the system to give similar results of that of processing tags and java scripts. We are going to categorise the identified URLs into three categories

Published by: The Mattingley Publishing Co., Inc.

which are Benign, Malicious and Spam websites. Benign websites are harmless and legitimates websites which are created to help or serve the users. Malicious websites are created with the intention of deceiving and stealing information from the user to use against them. Spam websites are basically adware website which bombard the users with ads which may further contain other malicious websites. The purpose of our system is to identify legitimate website from another harmful website and warn the user. The system prompts to enter a website name, once entered, gives result whether the website is benign, malicious or spam. According to anti-phishing reports, 1st quarter of 2014 saw the second highest number of phishing attacks ever [7]. The number of phishing attacks were 125,215. Therefore, we propose a machine learning approach where the system uses random forest and SVM classifiers and it does not take into consideration, the content of the websites. Due to rise in internet users throughout the world, more people are new to the concept of internet and well-versed attackers take advantage of their limited knowledge of the domain to trick them into stealing their information. Phishing attacks are very effective more than ever due to availability of cheap internet and the rise in users, it is necessary to protect the people from such attacks to give them a pleasant browsing experience online. This system does not effectively solve the problem of phishing attacks but a step in the right direction to develop and implement effective phishing detection techniques.

II. LITERATURE SURVEY

There are many techniques that are used to detect phishing attacks over the past years. Phishing attacks are broadly classified into two types: software-based techniques and user education-based techniques

Software-based techniques are further classified into heuristic based, visual similarity based and blacklist-based techniques.

User educational based approaches: Kumaraguru et al. [2] developed two training designs to help users to self-identify between phishing and non-phishing websites. Sheng.[1] created an educational interactive game known as “Anti-Phishing Phill” to educate users to protect themselves from common phishing attacks. Software-based techniques: This approach is further divided into 3 more categories.

A.Blacklist-based technique

This approach is basically a list is maintained which contains previously detected phishing websites. Whenever a website is checked for malicious features, it is referred with the blacklist which is usually updated in websites like phishtank.com. The major drawback of this approach Is that it cannot detect zero-day attacks. The list takes at least a few days to get updated with newer malicious websites and hence in those few days, financial loss can occur to the user. Sheng.[3] estimated that about 50%-80% of the phishing websites are added to blacklist after performing some kind of financial loss to users. Blacklist technique can also be very efficient and trust worthy since trusted moderators maintain a strict list of phishing websites, but this method can prove very ineffective in detecting zero-day attacks, zero-day attacks are those which occur on the day of release of the phishing website. Since this website is fairly new and it has not affected any users still, it is deemed as a harmless website. But users who access this website for the first time can be affected by such types of websites, even if they run a blacklist check, it won’t appear on the list as it has not been reported yet. This type of blacklist technique can affect users who access the website for the first time and it can also cause trust issues for even legitimate websites. Blacklist techniques are effective but still needs improvement in detecting zero-day attacks so such type of blacklist method is no longer used to confidently detect phishing websites. Better techniques to detect phishing attacks have been researched and implemented.

Published by: The Mattingley Publishing Co., Inc.

B.Heuristics-based technique:

In this technique, the heuristics design of the webpage is matched with the feature which commonly identifies a phishing website. This is almost similar to user based which includes education yourself to defend from phishing attacks. Malicious nature of the webpage by cross referencing the feature set.[4] Zero-day attacks (newly created web page for attack) can be identified using heuristic approach. Zhang.[5] proposed a method known as CANTINA which has rich set of features that can detect phishing websites. This heuristics-based method is a self-learning/ self-education technique where users are made to study the various common features that usually identifies a phishing website. By learning the various features, users can identify phishing websites to protect themselves from such attacks. The main disadvantage of the technique is that it is a tedious learning process and requires lots of time to learn the features which are very fast. This learning process cannot be used by normal users to protect themselves from such attacks. Attacks are usually used by companies to train their engineers to protect themselves from phishing attacks since company computers usually have very sensitive information and it is in the company's best interest to protect themselves from information theft. Lots of resources are also spent in heuristics-based method to buy the material to train the engineers. Time also is spent in training against phishing attacks which can be spent on other company projects which are the main goal of the company.

C.Visual-similarity techniques:

Visual similarity includes comparing the site with respective original sites to detect malicious features. This technique uses feature set like text content, HTML tags, CSS (cascading style sheet) features, image processing etc. Chen. [6] Proposed an anti-phishing approach by analysing the key visual features to detect a phishing website. This type of technique also is similar to the heuristic's method where certain features are studied and analysed in the website to detect its malicious nature. In this technique, visual features like website format, colour combinations used, placement of wordings, font of the letters, colour of certain features and distinguishing legitimate website features are analysed and searched for in the website to detect resemblance of phishing nature. Such a technique is also a self-learning technique where user himself identifies the phishing website. This technique includes certain drawbacks like users cannot always identify certain features to protect themselves, it is a self-learning process and hence it is very tedious. Some features can be very minute which the common user cannot identify and risks himself to accessing phishing websites. This technique can be very effective to users with who are frequent and long-time users of the internet, this technique can be very difficult to new and naïve users who have little experience with browsing on the internet. Since naïve users are the majority of the phishing attack victims, this cannot be considered as an effective technique to stop phishing attacks.

D.Use of Support Vector Machine (SVM)

Online phising classification uses the SVM with the theoretic game formatting techniques which has a prior knowledge function. This paper [8] new content-based feature extraction is done for the process of filtering. In this domain using the dynamic games of incomplete information in the oretic data mining framework is proposed in order to build the adversary- aware classifier for phising methods. Over the last years, phishing fraud through malicious email messages has been a serious threat that affects global security and economy, where traditional spam filtering technique shave shown to be ineffective.

E.Inference

Based on the literature survey done above, there exists no single method to detect all kinds of phishing attacks (websites, emails, etc.). The major

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8822 - 8830

drawback of blacklist technique is that it cannot detect zero-day attacks. We can use heuristics technique to identify zero-day phishing websites but fails to detect if embedded object presents in the webpage and false positive is also high in this approach. Visual similarity can detect the embedded object present in the webpages but it fails to detect zero-day attacks. Therefore, in this paper, we suggest a machine learning based anti-phishing technique that can detect a phishing website using URL (Uniform Resource Locator) features. This automated machine learning system can effectively detect phishing attacks of up to 90% accuracy which is very high compared to other user self-learning techniques

III. METHODOLOGY

The accuracy of phishing system to predict malicious nature depends upon the chosen feature set to distinguish between a phishing site and a malicious site. There can be a hundred features but we consider only 12 features which gives us the best results for a smaller feature range. These 12 features are specifically chosen which gives the most probability of chance to identify a phishing website. There exist hundreds of features but some feature occurs only once or twice in a dataset containing around thousand entries hence features are chosen such that they occur or commonly found in most of the phishing websites. Some features are also common in legitimate websites, such features are not considered to avoid confusion with legitimate websites

The below figure shows the ranking of most prominent features present in the phishing URLs.They are ranked based upon how common the features occur among the phishing URLs of the Dataset used.

```
Feature name : Importance
1 URL_Length   :  0.2031816343152776
2 web_traffic   :  0.1919786779548266
3 statistical_report   :  0.14143922576141502
4 age_domain   :  0.09070169369564164
5 Sub_domains   :  0.08373245421057791
6 dns_record   :  0.07782357721013357
7 domain_registration_length   :  0.068062296639371
8 tiny_url   :  0.06645095364580964
9 Prefix_suffix_separation   :  0.05429329875416499
10 Having_@_symbol   :  0.009027965689932072
11 Having_IP   :  0.006884889060233879
12 Redirection_//_symbol   :  0.005401601046016342
13 http_tokens   :  0.0010217320165998316
```

Fig. 3.1 Ranked list of features in phishing websites

A.FEATURE EXTRACTION

IP Address

A phisher may use IP address instead of the website name. This is done by the phisher to hide the identity of the website.

Sub domain

Phishing sites usually contains multiple sub domains in the URL. The domains are usually separated by a dot symbol (.). When a site contains more than 2 sub domains, the probability of it being a phishing website is high.

The usage of "@" Symbol in URL:

Phishing websites in common cases have "@" symbols in their URLs. This occurs a considerable amount of times, hence considered into the feature set.

The usage of "@" Symbol in URL:

To replicate the genuine nature of legitimate websites, phishing websites use dash (-) symbol to hide their malicious nature. This feature tries to exploit the brand value of other companies and it is a major feature in all phishing website URLs. This feature is common for about 50% of phishing website URLs. e.g.www.facebook-login-now.com.

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8822 - 8830

Length of URL:

Longer URLs is another trick to hide the phishing websites. The average length of URLs is 74 characters while 31% of phishing websites have URLs longer than 80 characters.

Suspicious keywords in URLs:

Phishing URLs usually use certain keywords to deceive the customer. Some keywords include login, pay, password, paypal, account, free, trusted etc. the probability of a phishing website having these keywords are very high.

Domain count in URL:

Phishing URLs commonly have multiple domains in the same URL. This is used to redirect the address to various other malicious websites.

Embedded domain in URL:

Another domain can be linked in the URL by using the "//" feature. When a URL has multiple links to various websites, it is classified as a phishing URL.

Main domain position in URL:

The position of main domain in the URL can say so much about the website. Phishing websites add branded companies after their domain to trick users into entering website.

Main domain position in URL:

The position of main domain in the URL can say so much about the website. Phishing websites add branded companies after their domain to trick users into entering website.

Example-http//:wxyz,facebook.com

Website Ranking:

For the sake of maintaining a consistent feature set, we only consider the top 100,000 websites in terms of usage as legitimate websites, any other website ranking above 100,000 chance of malicious content in the website is more. Although not all website over the 100000 rank are malicious, we consider this feature to maintain consistency in our features to obtain consistent results.

Age of website:

For our feature set, we consider websites which are less than 6 months old happen to have more chances of being a phishing website.

HTTPS protocol:

HTTPS is a protocol which is followed for a secure website usage. Most phishing website do not follow the HTTPS protocol. Only about 5% of phishing websites follow https protocol

The following features are observed as the most prominent features which are extracted among the 22 features of the original dataset. These features are ranked depending upon the number of times they occur among the phishing URLs of the

dataset. These features are used as the main criteria and tested with the multiple machine learning algorithms to identify the most suitable algorithm for prediction.

B.PROPOSED SYSTEM ARCHITECTURE

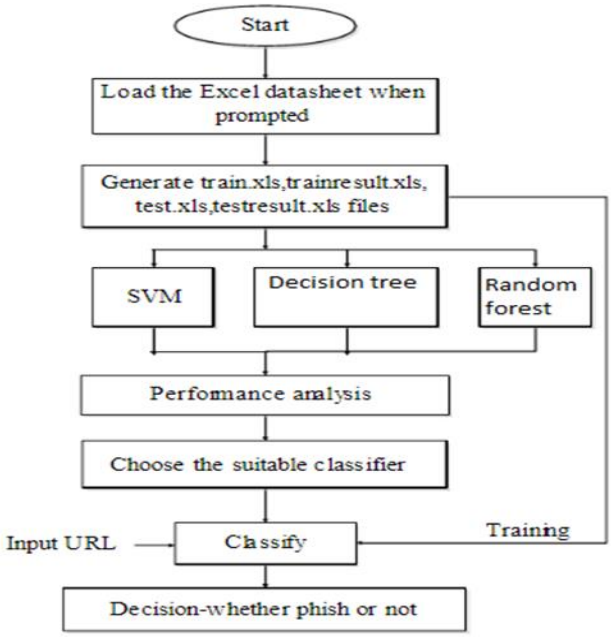

Fig. 3.2 Overview of proposed system

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8822 - 8830

The flow starts from collection of datasets for both training and testing from phisnet.com. Around thousand entries of URL are used for both training and testing datasets. The various hundreds of features are obtained that distinguish a phishing website. Features that are common for both legitimate and phishing websites are not considered due to cause of confusion. Feature wrapper is used to select only 12 features that occurs most commonly in phishing websites that distinguish them separately.

We use random forest algorithm to classify the dataset. The training of machine learning classifier is done using the extracted features. Finally, the evaluation is done by comparing the results with the training dataset. The result shows the accuracy of the system to detect categorize phishing websites when entered. The system consists of an UI prompting the user to enter a URL, once the URL is entered the system runs the URL in the phishing detection system and returns the result whether the URL is a benign website, malicious website or a spam website. This is the UI part of the system. We can observe the analysis on the Anaconda navigator which is a python distributor.

The datasets are tested against three data classifiers which are random forest, SVM algorithm and decision tree. Each dataset will have a certain algorithm which will give best results, we have to test multiple algorithms against the datasets to analyse the most suitable algorithm. In this project, we consider these three data classifiers as they have been proven to give best results for textual data as used in this project.

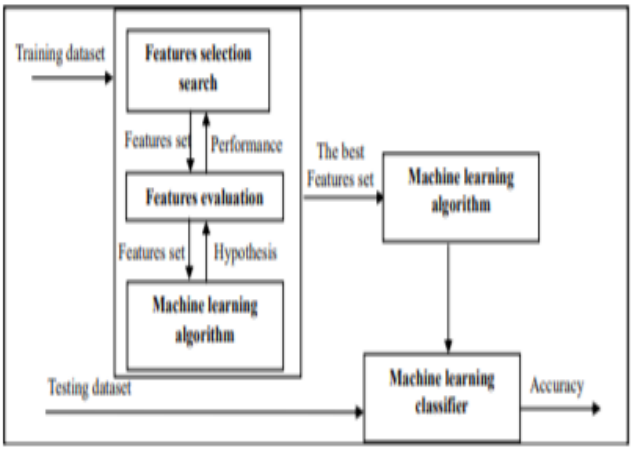

Fig. 3.3 Architecture diagram of Phishing System

The above figure explains the architecture diagram of the phishing detection system. The system first takes the training dataset which has around thousand entries of URL dataset to extract the major features that identifies a phishing website. Feature evaluation is done to select the most important features that are most common among phishing websites. This feature extraction is given to the machine learning algorithm, random forest classifier to detect phishing websites in the testing dataset. The best feature set is obtained from the feature evaluation and applied to the machine learning algorithm. Now, the training dataset is applied to the processed machine learning algorithm to produce the result in terms of accuracy. The end accuracy of the system can be further increased by increasing the number of entries in the URL phishing dataset and better feature evaluation to choose better features and choosing of better data classifier to increase the accuracy of the system. Accuracy is obtained by the evaluation matrix which used features such as true positive, true negative, false positive and false negative from the testing website which gives the accuracy.

The UI of the system is coded using python which produces a GUI that prompts the user to enter the website to be checked. The GUI receives the website and compares it with the machine learning algorithm and gives the result . Anaconda navigator which is a python distributor, which provides a desktop user

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8822 - 8830

interface to graphically manage datasets, Anaconda packages, environments and channels without using the command line commands. Anaconda navigator includes Jupyter, which is used on the chrome browser or any other browser for easy use of terminal Machine learning approach here is used to make use of the training dataset to obtain the features efficiently using self-learning and past examples and conditions to effectively choose options that are similar to current situations. The data classifier used is Random forest algorithm. It is a supervised, flexible and easy to use machine learning algorithm. It is used to classify the observed evaluation matrix which includes true positive, true negative, false positive and false negative to evaluate the matrix and gives the accuracy of the system.

IV. EXPERIMENTAL RESULTS

A.Evaluation metric

The evaluation involves completion of the matrix which has user's choice in particular, the above methods have been applied on the matrix to find the most efficient method through the process of cross validation.

TABLE I. Evaluation Metric

TN/True Negative	Legitimate websites are given as phishing
TP/True Positive	Legitimate websites are identified as legitimate
FN/False Negative	Phishing websites are identified as legitimate websites
FP/False Positive	Phishing websites are identified as such

B.Precision

Out of all the websites in the URL phishing dataset, how man websites were actually truly identified as phishing website.

It is given by:

$$\text{Precision}= \frac{TP}{FP+TP} \quad (1)$$

Accuracy of an algorithm is given:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

Here, TP represent the number of legitimate websites actually identified correctly as legitimate website, FP represents the amount of phishing websites wrongly identified as legitimate website, by using the precision formula, we can obtain the accuracy of the phishing detection system, When a system has 100% accuracy, it means it identifies phishing as phishing websites and legitimate as legitimates websites, all systems aim to obtain 100% accuracy.

TABLE 2: Comparison of algorithms

	TP	TN	FP	FN	Accuracy	Precision
Random forest	399	62	20	81	82.02%	95%
SVM	270	35	15	72	78.24%	94%
Decision tree	262	30	13	70	77.86%	95.27%

The above table shows the various observations of the three algorithms researched to best predict results of URL phishing, Random forest shows more accuracy, hence it is chosen as the algorithm to predict.

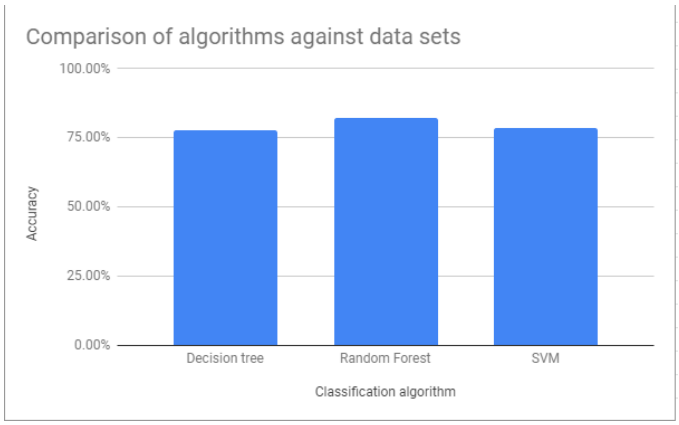

Fig. 4.1 Comparison of Algorithm accuracies

The above figure shows a bar graph of comparison between algorithm accuracies

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 8822 - 8830

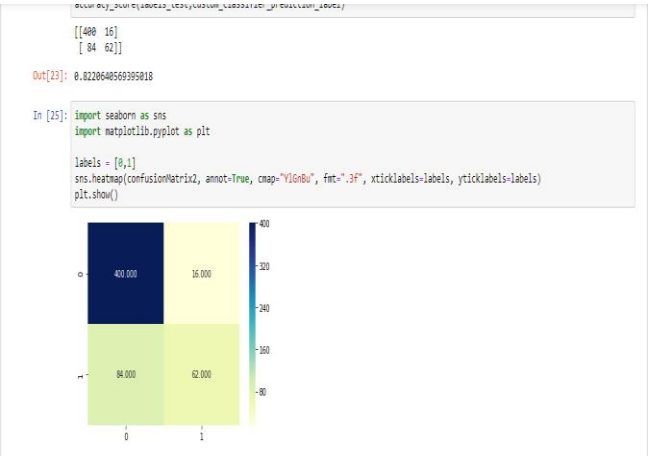

Fig 4.2 Accuracy of the system

The phishing dataset was obtained from phishnet.com for both the training and testing dataset. Around thousand entries of websites for both testing and training datasets were used. Total of 12 features were identified as the most important features that determine a website as a phishing website using feature evaluation. True positive rate (TP) and False positive (FP) is used to obtain the accuracy of the final phishing detection system which gives accuracy of 82% in the screenshot.

Then the random forest algorithm is then used as the primary machine learning algorithm in the system to predict future URL phishing entries.

V. CONCLUSION AND FUTURE WORKS

The system we proposed uses a machine learning approach to detect phishing websites. We have used 12 features to distinguish phishing websites from legitimate websites. We have obtained more than 90% accuracy while using Random forest data classifier to process the URLs. In future, more features can be added to further distinguish phishing and legitimate website and other data classifiers can be experimented to give higher accuracies of results. Better feature evaluation can give rise to better features that are most commonly found in phishing websites to better identify them. Larger datasets can also be used to increase the training dataset which can significantly increase the feature extraction in turns increases the accuracy of the system.

REFERENCES

[1]. Protecting people from phishing: the design and evaluation of an embedded training email system. In: CHI 2007 proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, pp 905–914. Kumaraguru P, Rhee Y, Acquisti A, Cranor LF, Hong J, Nunge E (2007)

[2]. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In: SOUPS 2007: proceedings of the 3rd symposium on usable privacy and security, ACM, New York, pp 8

[3]. 99 Sheng S, Magnien B, Kumaraguru P, Acquisti A, Cranor LF, Hong J, Nunge E (2007)

[4]. An empirical analysis of phishing blacklists. In: CEAS 2009 Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C (2009)

[5]. Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing E-mail. IJST 6(1):122–126.Almomani A, Gupta BB (2013)

[6]. Cantina: a content-based approach to detecting phishing web sites. In: Proceedings on WWW, ACM, New York, pp 639–648 Zhang Y, Hong JI, Cranor LF (2007)

[7]. Chen K-T, Huang C-R, Chen C-S (2010) Fighting phishing with discriminative key point features. IEEE Internet Community

[8]. APWG Q1 report, Available at:docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf (Last accessed on 6 November 2015)

[9]. Almomani A, Gupta BB, Atawneh S, Meulenberg A, Almomani E (2013) A survey of phishing email filtering techniques. IEEE CommunSurv Tutor 15(4):2070–2090

[10]. Anti Phishing Work Group (2014) Phishing attacks trends report. http://docs.apwg.org/reports/ apwg_trends_report_q2_2014.pdf

Published by: The Mattingley Publishing Co., Inc.

[11]. Kumaraguru P, Rhee Y, Acquisti A, Cranor LF, Hong J, Nunge E (2007) Protecting people from phishing: the design and evaluation of an embedded training email system. In: CHI 2007: proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, pp 905–914

[12]. Sheng S, Magnien B, Kumaraguru P, Acquisti A, Cranor LF, Hong J, Nunge E (2007) Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In: SOUPS 2007: proceedings of the 3rd symposium on usable privacy and security, ACM, New York, pp 88–99

[13]. Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C (2009) An empirical analysis of phishing blacklists. In: CEAS 2009

[14]. Almomani A, Gupta BB (2013) Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing E-mail. IJST 6(1):122–126

[15]. Zhang Y, Hong JI, Cranor LF (2007) Cantina: a content-based approach to detecting phishing web sites. In: Proceedings on WWW, ACM, New York, pp 639–648

[16]. Chen K-T, Huang C-R, Chen C-S (2010) Fighting phishing with discriminative key point features. IEEE Internet Community

[17]. Phishing URLs Dataset available at: https://www.phishtank.com