2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)

Phishing Emails Detection Using CS-SVM

Weina Niu, Xiaosong Zhang, Guowu Yang, Zhiyuan Ma, Zhongliu Zhuo
School of Computer Science and Engineering, and Center for Cyber Security
University of Electronic Science and Technology of China, UESTC
Chengdu, China
Email: niuweina1@126.com, {johnsonzxs,guowu}@uestc.edu.cn, yuliar3514@gmail.com, zhuozhongliu@126.com

***Abstract*—Phishing attacks are common online, which have resulted in financial losses through using either malware or social engineering. Thus, phishing email detection with high accuracy has been an issue of great interest. Machine learning-based detection methods, particularly Support Vector Machine (SVM), have been proved to be effective. However, the parameters of kernel method, whose default is that class numbers reciprocals in general, affect the classification accuracy of SVM. In order to improve the classification accuracy, this paper proposes a model, called Cuckoo Search SVM(CS-SVM). The CS-SVM extracts 23 features, which are used to construct the hybrid classifier. In the hybrid classifier, Cuckoo Search (CS) is integrated with SVM to optimize parameter selection of Radial Basis Function(RBF). Experiments are performed on a dataset consisting of 1,384 phishing emails and 20,071 non-phishing emails. Experimental results show that the proposed method has higher phishing email detection accuracy than SVM classifier with default parameter value. The CS-SVM classifier can obtain the highest accuracy of 99.52 percent.**

***Keywords*-APT; phishing email detection; SVM; Cuckoo search; RBF**

I. Introduction

In recent years, more sophisticated attacks have become a common problem in cyber security, one of which is Advanced Persistent Threat (APT) [1][2]. APT attackers often use social engineering techniques, for instance, phishing email, to invade the target network. Attackers send emails containing a phishing link to a malicious website or an attachment that contains malicious programs to target users. Then, attackers deceive target users to install a malicious program and then control the target host to steal sensitive information or cause damage. Additionally, the number of phishing emails continues to rise. Phishing Activity Trends Report [3] published by APWG finds 446,065 unique phishing sites, 18 million new malware samples in the second quarter of 2016 alone, and an average of more than 200,000 per day.

The content-based approach is the most accurate phishing detection method since it can identify new phishing attacks. Support Vector Machine (SVM) [4] is the most popular technology of all the content-based approaches. Machine learning-based technique has been shown to be effective to detect phishing email [5]. However, parameter selection of kernel method has a significant effect on the SVM classifier accuracy. Features also have an impact on identifying new phishing emails.

In this paper, we extract 23 features including body-based features, URLbased features, and header-based features to detect phishing emails. Then, we build a hybrid classifier based on these 23 features together, where Cuckoo Search (CS) [6] is used for parameter selection of kernel function. The hybrid classifier combining CS with SVM is evaluated on a testing dataset including old and new phishing emails and yields a better result than SVM classifier with default parameter of Radial Basis Function (RBF).

The remainder of the paper is organized as follows: Section 2 gives an overview of the related work on phishing email detection; Section 3 describes our proposed CS-SVM classification method and 23 features used for phishing identification; Section 4 introduces experimental setup and analyzes experimental results; Conclusion and future work are summarized in Section 5.

II. Related work

Some phishing email detection techniques have been proposed in recent years to reduce the damage caused by phishing attacks. Generally, phishing detection can be classified into the network-based approach, blacklist, whitelist, and content-based approach [7]. The network-based approach identifies phishing email through interfering TCP and UDP sessions. Since most of the message content are transmitted in encryption mode, the network-based approach is difficult to implement. Blacklist characteristic library of phishing email recognition. This would be the same for the whitelist. Although blacklist and whitelist are simple, they fail to detect new phishing attack. Moreover, blacklist and whitelist collection are time-consuming. The content-based approach is designed to obtain attack patterns, which has the highest detection accuracy among existing detection approaches. Meanwhile, the content-based detection approach often makes use of machine learning techniques to discover new phishing emails with high identification accuracy. The comparisons among these approaches are shown in Table I.

Features in the content-based detection literature are classified into URLbased and text-based. Text-based features include message headers and message body. Toolan and

0-7695-6329-5/17/31.00 ©2017 IEEE
DOI 10.1109/ISPA/IUCC.2017.00160

Table I
OVERVIEW OF PHISHING EMAIL DETECTION TECHNIQUES

Techniques	Advantages	Disadvantages
network-based	easy for blocking IP addresses	costly to implement and time-consuming
blacklist	easy	incapable of detecting new phishing attacks
whitelist	easy	high false positive rate
content-based	high accuracy	rely on high standard training datasets

Carthy [8] extracted 40 features including body-based feature, subject-based feature, URLbased feature, script-based feature, and senderbased feature to detect phishing. Their effectiveness was calculated through information gain and entropy. The experimental result showed that the most effective features were URL and script in an email containing URL. Huang et al. (2012) [9] proposed an SVM-based model. The model extracted 23 URL-based features, and the performance was evaluated. SVM-based approach produced a classification accuracy of 99.0 percent.

III. PROPOSED CLASSIFICATION METHOD

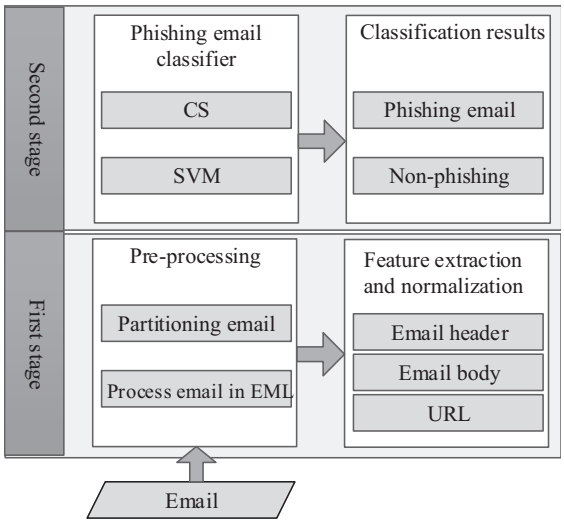

Fig. 1. The architecture of proposed method

In this section, we describe the proposed hybrid classifier for phishing emails detection. Our method focuses on feature extraction and classifier building. The architecture of the proposed method is shown in Fig. 1. The pre-processing phase includes: converting email to XML (Extensible Markup Language) format [**?**] and splitting email into three parts: header, body, and URL. The second phase is responsible for feature extraction based on header, body, and URL of email, and feature standardization. Then, we use CS-SVM as our classifier to identify phishing emails, where CS is used to optimize the parameter of kernel function in SVM.

A. Email Features Extraction

This section describes 23 features for detecting phishing emails.

1) Header-based Features: This section illustrates five header related features to detect phishing emails.

(1) Sending time of email is not in normal work time

Normal emails are often sent by many legitimate users in working time. On the contrary, phishers send malicious emails in the non-working time period to avoid drawing attention of detectors. Therefore, an email whose sending time is not in normal working time is a potential phishing email.

(2) Presence of dark copy

Email header in phishing emails may contain a dark copy. Hence, the presence of a dark copy in an email is checked and a binary value is recorded based on the presence or absence of the dark copy.

(3) Number of CCs

Phishers frequently send the same phishing emails to many different receivers. Thus, in this study, the total number of CCs in an email is extracted and used as a feature for classification.

(4) Whether a reply email

Phishing emails are not replies to the sender's emails. Thus, an email that is not a reply email is a potential phishing email.

(5) Presence of order, payment, RE- in the title

Title in phishing emails generally contains words like order, payment, RE- in order to deceive recipients. In this work, presence of order, payment, RE- in the title of an email is checked and a binary value is recorded for classifying phishing emails.

2) Body-based Features: In this work, body-based features include the following four features.

(1) Disparities between href attribute and LINK text

HTML anchor tag is used to establish a link to another website. If there is a disparity between the href attribute and the link text, then this link should be checked and if there is a disparity, then a positive of the binary value is recorded for detecting phishing emails.

(2) Presence of sensitive words in the LINK text

Link text in most phishing emails generally contains sensitive words, such as Click, Link, Here, Login, Update. Hence, the presence of sensitive words in all the LINK text in an email is checked.

(3) Presence of javascript

Javascript is generally used for hiding information from users. Thus, an email containing javascript string is a potential phishing email.

(4) Some specific verbs and You, Your in a sentence segment

If some specific verbs, like confirm, update, follow, access, click, enter, and protect, often appear in a sentence

segment of an email with You, or Your, this email is most likely to be a phishing email.

3) URL-based Features: This section gives explanations of 13 features related to URL in the email body.

(1) Presence of IP address in URLs

If an IP address is used as an alternative to the domain name in the URL of an email, such as http://15.9.31.123/fake.html, this email may be a phishing email. Therefore, in this work, we make use of the presence of IP address in URL to identify phishing emails.

(2) Number of URLs

Phishing emails frequently contain multiple URLs to illegitimate websites. Thus, the number of URLs is a feature for identifying phishing emails in this work.

(3) Number of dots in domain name is greater than 3

If the number of dots in a domain is greater than three, then the URL is classified as Phishing since it will have multiple subdomains. On the contrary, the legitimate domain name in the URL of an email has no subdomains, such as "http://baidu.com".

(4) Presence of "@" symbol in URLs

If an URL contains "@" symbol, the browser will ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

(5) Active duration of domain names in URLs is less than 6 months

This feature can be extracted from WHOIS database. Most phishing websites live for a short period of time.

(6) Length of URL is greater than 54 characters

Phishers make use of long URL to hide the doubtful part. Through the analysis of existing data sets, if the length of the URL is greater than 54 characters then an email containing this URL is classified as a phishing email.

(7) Presence of "-" character in URL

Legitimate email rarely uses dash symbol in URLs. However, phishers tend to use - character to deceive the recipient to believe that this is a legitimate web page.

(8) Presence of DNS record

Since the claimed identity of phishing URL is not recognized by the WHOIS database, the website is classified as Phishing when its DNS record is empty or not found.

(9) ALEXA ranking is greater than 100,000

Active duration of the malicious website is short, so the number of users and the number of pages that they visit is relatively small. On the contrary, the popularity of a legitimate website is high. The ALEXA value is a commonly used indicator of site traffic ranking statistics. Furthermore, in the worst scenario, legitimate websites ranked among the top 100,000. Thus, if ALEXA ranking of a domain name in an URL is greater than 100,000 or it is not found, then this website may be malicious.

(10) Registration duration is less than 2 months

Legitimate domains are regularly paid for several years in advance, for example the registration duration of google.com is nearly two decades.

(11) Number of http/https in an URL is greater than 1

The existence of "//" within the URL path means that the user will be redirected to another website. Thus, there are more than 1 http or https in a URL, then the URL may be redirected to a malicious website.

(12) Domain similarity measure

We find it interesting that many phishing domain names are similar to famous domain names. Thus, phishers not only make these domain names easy to remember but also make these domains more like normal ones. Therefore, we can make use of domain similarity measure to classify phishing emails.

(13) Using URL Shortening Services

URL shortening is used for making an URL considerably smaller in length and still leads to the required web page.

(14) Disparities between domain name server in URLs and domain name server of sender

Phishers often send malicious links whose servers reside in different countries or regions in order to hide the true attack source. Thus, domain name servers of URLs in phishing email do not reside in the same country of the attacker.

B. Classification

In this section, we introduce our hybrid classifier, which makes use of Cuckoo search to optimize parameter selection in the kernel function. CS is integrated with SVM to construct a hybrid classifier, CS-SVM, for selecting the optimum parameter in kernel function. In this paper, there are 23 features and several thousand emails in our testing environment. Thus, we select RBF to identify phishing email.

In the proposed CS-SVM algorithm, we select the traditional SVM algorithm as our fitness function. That is, this paper uses the value of to generate the hyperplane which minimizes the training errors and also maximizes the margin with the correctly classified data points. Then, this work calculates the classification error about normal emails and phishing emails using the current hyperplane. We modify the value of according to Cuckoo Search algorithm until the classification error remains unchanged or the maximal number of CS iterations reaches. The detailed process of CS-SVM is illustrated in Algorithm 1.

IV. Experimental Results and Analysis

A. Evaluation Metrics

There are two phases in supervised machine learning, namely the training phase and testing phase. To evaluate the performance of the proposed classifier, this paper chooses three commonly used evaluation metrics, which are True Positive Rate (TPR), False Positive Rate (FPR), and accuracy [10]. TPR and TFR are explained as follows: the proportion of correct identified phishing emails and the sum

Algorithm 1 CS-SVM algorithm

Require: NR the number of runs, NC the number of Cuckoos, D the dimension of objective search space, Θ the tuning parameter value range of γ in RBF, p_∂ the probability of finding eggs of exotic Cuckoo, $T = (x_i, y_i), (i = 1, ..., N; y_i = -1 | y_i = 1)$ training email samples with two classes

Ensure: γ parameter value which minimizes the classification error

1) Initialize nests location $\gamma_{g,i} (i = 1, ..., NC, g = 1)$
2) Generate the hyper planes $SVM_{g,n}$
3) Calculate classification error $E_{g,n}$ using $SVM_{g,n}$
4) Select the best solution $\gamma_{g,i}$ with the minimum error
5) **while** $g \leq NR$ **do**
6) Generate new location $\gamma_{g+1,i}$ using Levy flight
7) Generate the hyper planes $SVM_{g+1,n}$
8) Calculate new classification error $E_{g+1,i}$ of $\gamma_{g+1,i}$
9) Select the best solution $\gamma_{g+1,i}$ with the minimum error
10) **if** $E_{g+1,i} > E_{g,j}$ **then**
11) Retain the best solution $E_{g+1,k}$
12) Discard other solutions according to p_∂
13) Generate new values to replace discarded values using stochastic preference swimming
14) **end if**
15) Run time add 1
16) **end while**
17) Output γ

of phishing emails classified by SVM, the proportion of correct identified non-phishing emails and the sum of non-phishing emails classified by SVM. Precision, recall, and accuracy are defined as follows.

$$\begin{aligned} Precision &= \frac{N_{TP}}{N_{TP} + N_{FP}} \\ Recall &= \frac{N_{TP}}{N_{TP} + N_{FN}} \\ Accuracy &= \frac{N_{TP} + N_{TN}}{N_P + N_F} \end{aligned} \tag{1}$$

where, N indicates the number, N_P represents the number of actual phishing emails, N_F represents the number of actual non-phishing emails.

B. Experimental Setup

The experimental data used in this paper is collected from three data set and consists of phishing emails as well as non-phishing emails. The phishing email set consists of two different corpora, in which one is 1,203 old phishing emails collected by Jose Nazario in 2005 [11] and the other is 181 up-to-date phishing emails reported in MillerSmiles archive [12]. The 20,071 non-phishing emails are collected from the public Enron email set [13] by the CALO Project consisting of more than 150 employees.

Table II
THE DISTRIBUTIONS OF TRAINING SET AND TESTING SET

Data set	P-e	S-P-e	N-P-e
training	50%∼90%	0	50%∼90%
testing	10%-181∼50%-181	all	10%∼50%
Notes: P-e, phishing email; S-P-e, state-of-art phishing email; N-P-e, non-phishing email			

In order to evaluate the true positive rates and false positive rates of our proposed optimized SVM classifier, we conduct an evaluating experiment to select the optimal parameter in our training data set including part of non-phishing emails and phishing emails from the Nazario corpora. The distributions of training set and testing set is shown in Table II.

In our CS-SVM algorithm, there are several parameters requiring initialization. The parameter settings in experiment are shown in Table III.

Table III
PARAMETERS SETTING IN EXPERIMENT

Parameter	Description	Value
n	number of Cuckoos	25
A	minimum value of γ	0
B	maximum value of γ	1
dimension	dimension of input value	15
iteration	number of iteration	1000
p_∂	probability of selecting new nests	0.25

C. Experimental Analysis

In reference to different γ, different email data and different features, a comparison is made to verify the effectiveness of our proposed hybrid classifier.

1) Identification Accuracy With Different γ: Experiments are performed to evaluate phishing email identification accuracy at different parameters in kernel function. We first select sixty percent of non-phishing emails from Nazario corpora and the same proportion of phishing emails from Enron randomly. Identification accuracy changes with different parameter γ, which is shown in Fig. 2. SVM yields to the highest classification accuracy of 99.5064 percent when the value of γ is 0.77. Also, the number of emails that are classified correctly keeps invariable with γ increases. However, the identification accuracy of SVM reduces with γ decreases. If the default parameter is set to be greater than 0.77, we can obtain the optimal SVM classifier in the current email data. However, when γ takes the default value 0.5, the phishing emails identification rate is less than our method. Also, there are many local optimal points in the range of 0 to 1, such as 0.03, 0.13, 0.23, 0.29, and 0.48.

We randomly select sixty percent and seventy percent of emails from Nazario corpora, Enron set, respectively. The phishing identification precision of SVM increases with γ , based on the graph of false positives and false negatives

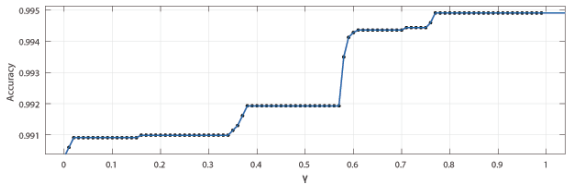

Fig. 2. Identification accuracy with different parameter

change is shown in Fig. 3. The false positive rate remains about the same along with γ changes. Decreased parameter results in discrimination of the false negative rate. Here, the classifier has lowest false positive rate and false negative rate when γ is 0.79 and 0.62, respectively. Moreover, the false negative rate can seriously affect the performance of classifier. In different training data sets, different γ values make SVM yield the highest classification accuracy. Thus, parameter in kernel function is set to default value does not meet the real situation.

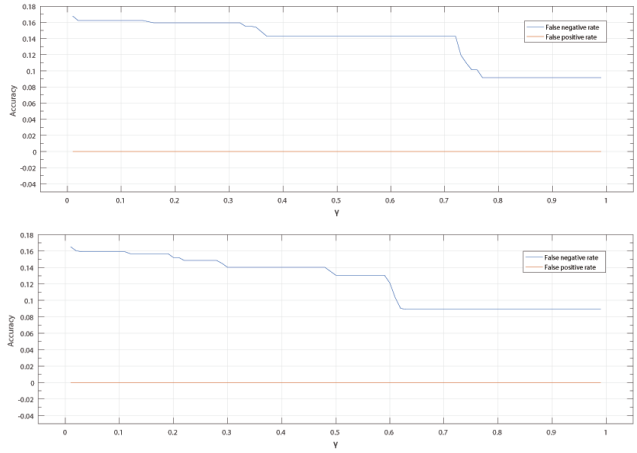

Fig. 3. False positives and false negatives with different parameter

2) Identification Accuracy Under Different Email Data: The performance of CS-SVM is evaluated by implementing some experiments. Experimental results of our proposed CS-SVM and SVM that using a default . As shown in Table IV, CS-SVM yielded to a higher phishing email identification rate than SVM classifier. Under different testing proportions, our proposed CS-SVM algorithm selects different parameters to maximize phishing email identification accuracy. Moreover, the optimal parameter value is different. Also, the ratio of correctly classified non-phishing emails in two different algorithms is the same and remains unchanged with different testing ration. The reason is that phishing emails often change their behavioral characteristics to evade being detected.

As shown in Fig. 4, when the ration of training set and testing set is one to one, CS-SVM and standard SVM yielded

Table IV
CLASSIFICATION RESULTS OF CS-SVM AND SVM

R	Algorithm	γ	TP	FN	TN	FP
1:1	CS-SVM	0.79	726	56	10035	0
	SVM	0.5	696	86	10035	0
5:6	CS-SVM	0.64	840	62	12042	0
	SVM	0.5	806	96	12042	0
5:7	CS-SVM	0.79	946	77	14049	0
	SVM	0.5	909	114	14049	0
5:8	CS-SVM	0.92	1059	84	16056	0
	SVM	0.5	1025	118	16056	0
5:9	CS-SVM	0.57	1167	96	18063	0
	SVM	0.5	1155	108	18063	0
1:2	CS-SVM	0.74	1277	107	20071	0
	SVM	0.5	1218	166	20071	0

a TPR of 92.8% and 89%, respectively. The TPR of CS-SVM is higher than that of SVM. Also, phishing email identification accuracy of CS-SVM is better under different testing rations. Moreover, the highest of TPR is 93.12%, which is about four percent higher than SVM with a default value. However, the FPR of two different classifiers is the same. These results revealed that the overall performance of CS-SVM outperformed traditional SVM.

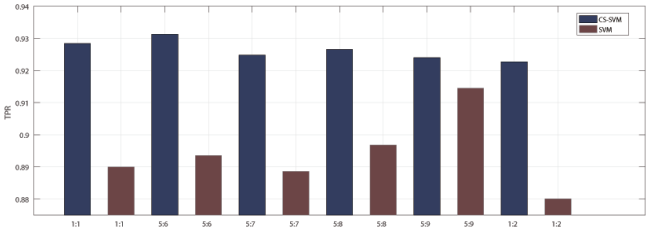

Fig. 4. TPR and FPR with different classifier

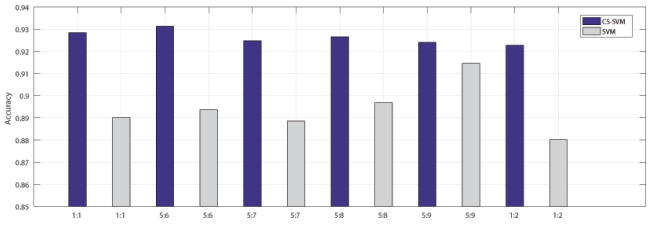

Fig. 5. Accuracy with different classifier

As shown in Fig. 5, CS-SVM was compared with traditional SVM classifier and had a classification accuracy of more than 91 percent. In the beginning, phishing emails detection rate increases with the ratio of testing. However, classification accuracy decreases when the ratio of training test and testing rate are less than 5:6. This is because the amount of new phishing samples appear in the testing data.

V. Conclusion And Future Work

Phishing email, especially spear phishing, has brought serious challenges to cyber security. The best current phishing email detection method is based on content and support vector machine. Since email features are far less than training emails detection, Radial basis function kernel with default parameter is used to deal with phishing email identification. However, default parameter cannot make SVM classifier perfect. In this work, we combine optimal characteristics of Cuckoo search algorithm with SVM classifier to select optimal parameter value in kernel function. In addition, we select 23 features including header, URL, and body. We perform experiments on a dataset, which contains three archives. Experimental results show that CS-SVM has a higher phishing email detection accuracy at different training set. This indicates that the proposed method is better than SVM classifier with default parameter value.

The future work will be devoted to the optimization of CS-SVM because our proposed method is performed on a single machine. Thus, we hope that our optimization algorithm could be run on a distributed platform. For example, we can use map-reduce to calculate fitness function of different γ in parallel.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61572115), the Key Basic Research of Sichuan Province (Grant No. 2016JY0007).

References

[1] Granadillo, Gustavo Gonzalez and Garcia-Alfaro, Joaquin and Debar, Hervé and Ponchel, Christophe and Martin, Laura Rodriguez Considering technical and financial impact in the selection of security countermeasures against Advanced Persistent Threats (APTs), *New Technologies, Mobility and Security (NTMS), 2015 7th International Conference on*, IEEE, 2015, pp. 1–6.

[2] Chandra, J Vijaya and Challa, Narasimham and Pasupuleti, Sai Kiran, "A practical approach to E-mail spam filters to protect data from advanced persistent threat," *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on*, IEEE, 2016, pp. 1–5.

[3] "APWG Phishing trends reports," http://www.antiphishing.org/.

[4] Basnet, Ram and Mukkamala, Srinivas and Sung, Andrew H, "Detection of phishing attacks: A machine learning approach," *Soft Computing Applications in Industry*, 2008, pp. 373–383.

[5] Bergholz, Andre and Chang, Jeong Ho and Paass, Gerhard and Reichartz, Frank and Strobel, Siehyun, "Improved Phishing Detection using Model-Based Features," *CEAS*, 2008.

[6] Kaveh, A, "Cuckoo search optimization," *Advances in Metaheuristic Algorithms for Optimal Design of Structures*, 2017, pp. 321–352.

[7] Gupta, BB and Tewari, Aakanksha and Jain, Ankit Kumar and Agrawal, Dharma P, "Fighting against phishing attacks: state of the art and future challenges," *Neural Computing and Applications*, 2016, pp. 1–26.

[8] Toolan, Fergus and Carthy, Joe, "Feature selection for spam and phishing detection," *eCrime Researchers Summit (eCrime)*, 2010, pp. 1–12.

[9] Huang, Huajun and Qian, Liang and Wang, Yaojun, "A SVM-based technique to detect phishing URLs," *Information Technology Journal*, 2012, vol. 11, no. 7, p. 921.

[10] Adewumi, Oluyinka Aderemi and Akinyelu, Ayobami Andronicus, "A hybrid firefly and support vector machine classifier for phishing email detection," *Kybernetes*, 2016, vol. 45, no. 6, pp. 977–994.

[11] [Nazario, J, "The online phishing corpus", http://monkey.org/~jose/wiki/doku.php.

[12] "The online phishing corpus," http://www.millersmiles.co.uk/.

[13] "Cohen WW (2016) Enron email dataset," https://www.cs.cmu.edu/*./enron/.