Decision Support Systems 88 (2016) 18–27

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder

Choon Lin Tan[a], Kang Leng Chiew[a,*], KokSheik Wong[b], San Nah Sze[a]

[a]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[b]Faculty of Computer Science and Information Technology,University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form 28 March 2016
Accepted 22 May 2016
Available online 1 June 2016

Keywords:
Phishing detection
Identity keywords
N-gram
Weighted URL tokens
Search engine

ABSTRACT

This paper proposes a phishing detection technique based on the difference between the target and actual identities of a webpage. The proposed phishing detection approach, called PhishWHO, can be divided into three phases. The first phase extracts identity keywords from the textual contents of the website, where a novel weighted URL tokens system based on the N-gram model is proposed. The second phase finds the target domain name by using a search engine, and the target domain name is selected based on identity-relevant features. In the final phase, a 3-tier identity matching system is proposed to determine the legitimacy of the query webpage. The overall experimental results suggest that the proposed system outperforms the conventional phishing detection methods considered.

1. Introduction

In this modern age of information technology, consumers are dealing with more products and services through the online channel. Therefore, having multiple online accounts (e.g., email account, banking account, social networking account) have become a norm for most people. This technological trend is exposing internet users to a rising threat of online identity theft known as phishing [17].

Phishing websites are counterfeit websites designed to deceive victims and steal their account login credentials, credit card numbers or other personal secrets. Phishers usually entice victims to the phishing website by sending emails containing the fraudulent URL and some threatening messages such as possible account termination, and fake alert on illegal transaction [9]. At the phishing website, the phishers will capture sensitive information submitted by the victims.

The severity of phishing threats in recent years continues to escalate, based on statistics gathered from security organizations. For instance, a total of 42,212 unique phishing websites was reported in June 2014 by the Anti-Phishing Working Group [2], whereas the financial loss inflicted upon worldwide organization in December 2014 was estimated to be $453 million [10]. These alarming trends have resulted in the loss of consumers' trust in using E-commerce websites because they are feared to become fraud victims [6]. In summary, phishing attacks have resulted in widespread leakage of sensitive information, monetary loss and crippled businesses' reputation.

The key factor that makes phishing possible is the human behaviour when interacting with electronic communication channels. Dhamija et al. [8] identified several user tendencies that are exploited by phishing attacks. For instance, a typical user is often unaware of the significance of common security indicators such as the Secure Sockets Layer (SSL) icon and digital certificate on the browser address bar. As a result, these useful indicators are often ignored. In addition, some users are confused on how a legitimate URL is supposed to resemble, thus they rely on the webpage contents to determine its genuineness [18]. A recent assessment by Alsharnouby et al. [1] reveals that participants with phishing awareness can only achieve 53% of average success rate in identifying phishing websites. These studies have proven that both normal and technical users can be easily deceived by phishing webpages. Hence, it is crucial to have an efficient phishing detection system, where users can be effectively safeguarded from phishing attacks.

To compensate for the human limitations in detecting phishing websites, automated solutions have been introduced in conventional web browsers and security applications. Most solutions rely on blacklists (e.g., Google Safe Browsing list, PhishTank list) that

[*] Corresponding author. Tel.: +6082 58 3762.
E-mail addresses: colin89lin@gmail.com (C. Tan), klchiew@unimas.my (K. Chiew), koksheik@um.edu.my (K. Wong), snsze@unimas.my (S. Sze).

http://dx.doi.org/10.1016/j.dss.2016.05.005
0167-9236/© 2016 Elsevier B.V. All rights reserved.

are compiled from automated link analysis or manual submission by volunteers. Zhang et al. [24] and Sheng et al. [21] tested several browser-based phishing detection tools, and concluded that most tools fail to detect phishing websites that are yet to be added into the blacklist or otherwise known as zero-day phishing websites. Recently, Purkait [19] tested the latest commercial browsers and security applications, where 12 out of 14 tools have failed to detect any of the phishing websites that existed for only a few minutes old. In short, the related studies in [19, 21, 24] have raised critical concerns on the reliability of the conventional tools. Furthermore, the existing tools also face challenges from increasingly sophisticated phishing threats, such as phishing webpages injected into existing legitimate websites and cloning of legitimate webpages using phishing toolkit [3]. Despite the existence of conventional phishing detection tools, the general public remains exposed to high risk of becoming phishing victims.

The detection of phishing webpages is extremely important, since it is the core mechanism for the mitigation of phishing attacks. When the detection phase is functional and accurate, appropriate warnings can be issued and actions can be applied to protect the victim or minimise the effects of the phishing attack. As such, we propose PhishWHO in this paper, which is an extension of our proposed phishing detection system in [22]. PhishWHO is based on a permanent phishing characteristic that stays intact over time, namely the discrepancies between the target and the actual identities of the query webpage. Here, target identity is defined as the domain name belonging to a legitimate brand that the phishing webpage deceptively represents, while actual identity refers to the query webpage's domain name. For legitimate webpages, the target identity often point to its own domain name, while phishing webpage does not. As such, the proposed method checks whether the query webpage is promoting itself, or promoting another existing legitimate webpage. By applying the proposed information processing techniques, both the target identity and the actual identity can be systematically derived from the query webpage. Hereinafter, the term "identity" and "domain name" shall be used interchangeably.

Several other enhancements are also incorporated in this paper, which distinguish it from our previous work in [22]. Our contributions include: (a) Exploiting the differences between the target and actual identities of a webpage to detect zero-day phishing webpages; (b) proposing a novel weighted URL tokens system based on N-gram model to overcome language limitations in phishing webpage detection; (c) exploiting indirect identity relationships to reduce false positives, and; (d) offering long-term effectiveness by leveraging on permanent phishing characteristic.

The remainder of this paper is organised as follows: Section 2 briefly reviews the scholarly works related to this research. Section 3 puts forward the proposed method. Section 4 describes the experiment setup and summarises the results. Section 5 discusses the merits and limitations of the proposed method. Finally, Section 6 concludes this paper.

2. Related works

A broad range of automated anti-phishing techniques have been introduced over the years. This section reviews the conventional anti-phishing techniques available.

2.1. Text-based detection

Zhang et al. [25] proposed CANTINA, a text-based phishing detection technique that extracts keywords from a webpage using the term frequency-inverse document frequency (TF-IDF) algorithm. The keywords are searched on Google, where the query webpage will be classified as legitimate if its domain name exists among the search results. Enhancements to CANTINA are proposed in [13], including an improved HTML parsing method and text-handling. The main problem is the reliance of TF-IDF on language-specific word list, thus Zhang et al. [25] and Komiyama et al. [13] are ineffective in classifying non-English webpages. Similar language limitation is found in [20].

2.2. Visual-based detection

Fu et al. [11] proposed using the Earth Mover's Distance (EMD) to assess visual similarities between suspected webpages and legitimate webpages. To calculate visual similarities, the webpages are sampled into low resolution images and represented by image features, i.e., dominant colour category and corresponding centroid coordinate. However, phishers can evade their method by making phishing webpages that appears less similar with the legitimate webpages. In another work, the same EMD algorithm is combined with a text classifier based on the naive Bayes rules to detect phishing webpages [23]. The bottleneck in visual-based techniques is the need for a large image database of legitimate websites. Since it is costly to maintain an up-to-date image database, visual-based techniques are impractical for widespread adoption. Chiew et al. [7] attempts to address this weakness by proposing an approach to extract logo from the webpage and submit it to Google image search, where the target identity can be determined.

2.3. Feature-based detection

Several researchers have shown that URL features are viable in determining whether a website is a phishing website. Le et al. [14] proposed an approach based on URL features such as length of full URL, domain name, directory, file and the number of symbols. These features are then fed into a classifier to compute the phishing probability. Another subset of feature-based techniques detects phishing based on search result features. For instance, Huh and Kim [12] queried Google, Yahoo, and Bing with the query webpage URL to obtain the number of search results and website ranking. These features are forwarded to a K-Nearest Neighbour (KNN) based classifier to compute the legitimacy of a webpage. Although feature-based techniques has the advantage of detecting zero-day phishing webpages, they often suffer from high false positive rates.

2.4. Identity-based detection

Liu et al. [15] focus on webpage identity analysis to detect phishing, where the Semantic Link Network (SLN) is proposed. SLN consists of weighted paths linking a set of webpages associated with the query webpage. Several metrics of the SLN are calculated to find the target identity and determine the legitimacy of the query webpage. Their proposed method are further improved in [16], resulting in the replacement of the SLN module by a parasitic community module. Motivated by [16], a similar strategy is proposed by [20] to identify the target domain name. However, [15, 16, 20] may fail to find the target domain name on phishing webpages that do not contain any URL belonging to the targeted legitimate website, thus resulting in false negative detections. In other words, phishers can easily bypass these techniques by hosting all the webpage resources (e.g., images, JavaScript, CSS) on the same phishing website.

2.5. Prevention-based techniques

Prevention-based techniques such as password management tools and 2-factor authentication focus on distorting the phishing infrastructure to protect the user's credentials [5]. For instance, a password management tool may add IP address string to the password whenever the user creates a new account at a legitimate website. When the user tries to log into a phishing website, a wrong

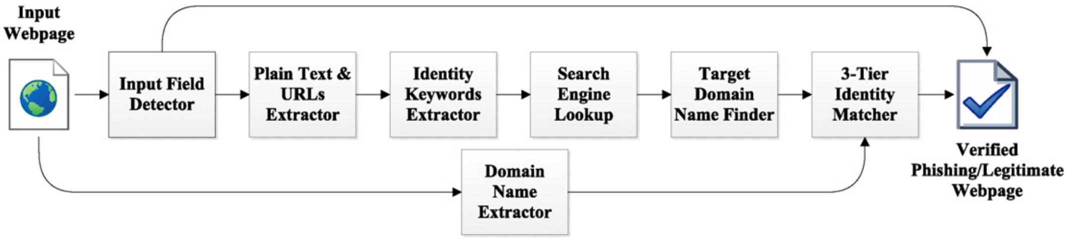
Input Webpage
Input Field Detector
Plain Text & URLs Extractor
Identity Keywords Extractor
Search Engine Lookup
Target Domain Name Finder
3-Tier Identity Matcher
Domain Name Extractor
Verified Phishing/Legitimate Webpage

Fig. 1. The architecture of PhishWHO.

password will be constructed based on IP address string of the phishing website, thus preventing the real password of the user from being submitted. On the other hand, 2-factor authentication make use of a hardware that is possessed only by the genuine user to perform authentication. For example, a One Time Password (OTP) is often send to the mobile phone of the customer by Short Message Service (SMS). To complete the transaction, the customer must supply the OTP within several minutes before it becomes invalid.

3. Proposed method — PhishWHO

The proposed method focuses on the detection of phishing webpage that has been loaded on the client's browser. The main components of the proposed method are shown in Fig. 1.

3.1. Input field detector

The first step of the proposed method finds the existence of inputs fields where victims' credentials may be leaked to the phisher. Note that all input types are taken into consideration. When no input fields are detected, the webpage can be safely declared as harmless. However, if at least one input field is found, the process continues to the remaining steps of the proposed method.

3.2. Plain text and URL extraction

Plain texts are extracted from several identity-relevant tags in the HTML source code such as the meta tags, title tag, body tag and *alt* attribute of all tags. On the other hand, URLs are retrieved from the *src* and *href* attributes of tags within the HTML source code.

3.3. Identity keywords extraction

After the plain text and URLs are retrieved from the webpage, PhishWHO proceeds to extract identity keywords from the plain text. Most conventional approaches employ the TF-IDF algorithm to extract keywords. To overcome the limitation of TF-IDF discussed in Section 2, we introduce a novel algorithm called the weighted URL tokens system.

3.3.1. Weighted URL tokens system

The weighted URL tokens system is a core component of the identity keywords extractor that functions as a robust weight generator. For each word in the plain text, the weighted URL tokens system calculates several metrics based on the structure of URLs extracted in the previous step.

To formulate the concept of the weighted URL tokens system, we specifically considered the visual perception of users and explored the question of what makes users convinced that the website they visit is a trustworthy or genuine one. Fig. 2 shows the common appearance of a URL in the web browser. It is evident that words appearing nearer to the left hand side (LHS) of the URL are more likely to capture the attention of users. Phishers exploit this visual appearance by intentionally placing identity keywords towards the LHS of the URL to mimic the appearance of legitimate URLs. This finding is also discussed by [8]. Based on this, a word is assumed to be more important when it appears on the LHS. This forms the basis of the proposed weighted URL tokens system.

Let the extracted tokens of the plain texts and URL be denoted as T_{plain} and T_{url}, respectively. If T_{plain} is a substring in T_{url}, a weight will be assigned to T_{plain}. The process begins by segmenting the URLs into tokens delimited by the forward slash. Taking the URL https://www.paypal.com/ma/cgi-bin/webscr?cmd=_registration-run&from=PayPal as an example, the segmentation results are shown in Table 1.

Next, the importance level shown in Fig. 2 is quantified. Given the i-th distinct word in the HTML textual content, its final weight, denoted by W_i, is calculated as Eq. (1).

$$W_i = \frac{l_i}{n} \sum_{k=1}^{L} \frac{N_k}{k^2}, \tag{1}$$

where l_i denotes the string length of the i-th distinct word, n represents the total number of URLs extracted from the webpage, k is the index of the URL token where the i-th word occurs, L is the index of the last URL token, and N_k is the total number of occurrence of the i-th distinct word in URL tokens of index k.

The rationale of Eq. (1) is further explained. When a webpage contains a large amount of URLs, the number of occurrences of a distinct word across all the URLs may increase, which in turn will increase the overall weight. Hence, the usage of n, i.e., the total number of URLs, as a denominator in Eq. (1) is intended to stabilise the overall weight when processing webpages with varying amount of URLs. The expression k^2 is intended to quantify the decreasing importance level for URL tokens that appear nearer to the RHS (i.e., having larger k) as shown in Fig. 2. In addition, k^2 also prevents the overall weight from becoming too large, since k is limited while N_k may be potentially big, which may overweight the overall weight.

Occasionally, the legitimate webpages may contain URLs of associated domains. Note that these URLs are not ignored by the proposed method, but they are processed normally as a subset of all the extracted URLs. If a T_{plain} occurs in any of the URLs, i.e., URLs of associated domains or URLs of the legitimate domain, W_i will be assigned to the T_{plain} based on Eq. (1). If a T_{plain} does not occur in any of the URLs, its W_i will be set to 0. A T_{plain} having $W_i = 0$ essentially means that it is unlikely to be an identity keyword.

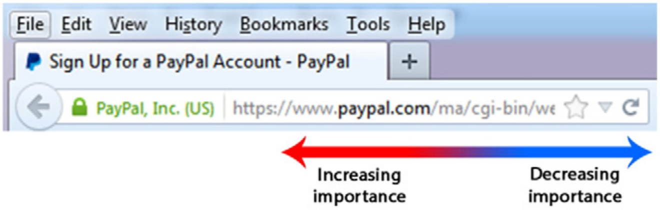
File Edit View History Bookmarks Tools Help
Sign Up for a PayPal Account - PayPal
+
PayPal, Inc. (US)
https://www.paypal.com/ma/cgi-bin/we
Increasing importance
Decreasing importance

Fig. 2. Perception of users when looking at URL in web browser.

Table 1
Index of URL tokens.

Index	URL tokens, T_{url}
–	https:
1	//www.paypal.com
2	/ma
3	/cgi-bin
4	/webscr?cmd=_registration-run&from=PayPal

When W_i has been computed for all T_{plain}, five T_{plain} with the highest W_i are selected and placed in set X. Subsequently, words in X are eliminated, one word at a time, by applying a sequential rule-based selection method which favours T_{plain} that have the characteristics of identity keyword. The sequential rule-based selection method is defined as follows:

- If the current word is a domain name, discard all other words in X that exist as substrings of the current word.
- If the current word contains at least one uppercase letter, discard all other words in X that contain no uppercase letters.

If X becomes an empty set after applying the sequential rule-based selection method, T_{plain} with the highest W_i is restored to set X. The target identity keywords are then defined as $K = X$.

The proposed weighted URL tokens system is straightforward in extracting identity keywords which consist of a single word (i.e., uni-gram). However, some organizations have names that are made up from several words, such as Bank of America, Hong Leong Bank, Commonwealth Bank of Australia, Sports Illustrated, and Capital One. This may lead to incomplete identity keywords being extracted and propagated to the remaining components in the proposed method. To solve the problem of multi-gram identity keywords extraction, we propose to extend the weighted URL tokens system as shown in Fig. 3.

3.3.2. N-gram model

Previous works on anti-phishing mainly focused on how to extract uni-gram identity keywords, while little attention was put on the extraction of multi-gram identity keywords. Therefore, our pioneering work in this section is significant for identity-based phishing detection.

Here, the N-gram models can be described as shifting a window across a string of words, so that only N words are considered at one time, and the uni-gram model is applied. As an example, Table 2 shows the results of applying the N-gram model on the sentence "*Bank of America is a well-known financial institution*" for $1 \leq N \leq 5$. The maximum value of N is determined empirically based on our observation that multi-gram identity keywords rarely exceeds $N = 5$.

The first step in applying the N-gram model is to pre-process the N-grams to increase the likelihood of striking a match with the URL substrings or tokens. The pre-processing steps are defined after analysing the common patterns of website domain names such as the utilisation of acronym, removal of whitespace, and replacement of whitespace with hyphen.

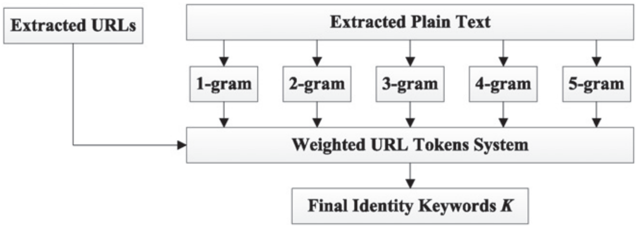

Fig. 3. Internal architecture of identity keywords extractor.

Table 2
N-grams for a sample sentence.

N	N-grams
1	"Bank", "of", "America", "is", "a", "well-known", "financial", "institution"
2	"Bank of", "of America", "America is", "is a", "a well-known", "well-known financial", "financial institution"
3	"Bank of America", "of America is", "America is a", "is a well-known", "a well-known financial", "well-known financial institution"
4	"Bank of America is", "of America is a", "America is a well-known", "is a well-known financial", "a well-known financial institution"
5	"Bank of America is a", "of America is a well-known", "America is a well-known financial", "is a well-known financial institution"

An acronym is the concatenation of the first letter of every word in a set of multi-gram identity keywords. Many organizations are known to use acronym in their website domain name, such as Internal Revenue Service (irs.gov), United Services Automobile Association (usaa.com), etc. Therefore, the acronym for every N-gram is computed to be used later in the matching process. All whitespace present in every N-gram are also removed, since there is no whitespace in a valid domain name. A minority of domain names uses hyphen to combine words together. To account for this, hyphens are removed in the pre-processing as well. Some sample results of N-gram pre-processing are shown in Table 3, where the original N-gram, concatenated N-gram and N-gram acronym are denoted as N_o, N_c and N_a, respectively.

Before the N-grams are ready for the weighted URL tokens system, filtering is performed to eliminate noisy data. Every N-gram is analysed to find the number of uppercase letters and the number of alphabets. If zero uppercase letters are found or the N-gram is fully numeric, we shall discard it and proceed to the next N-gram. These filtering steps are important to verify that the N-gram has some basic characteristics of identity keywords. It is important to note that these filtering steps are not ambiguous with the final step of the sequential rule-based selection method, since they are used for two different purposes. The sequential rule-based selection method is used specifically for selecting identity keywords after W_i has been computed for all T_{plain} of the uni-gram. On the other hand, the N-gram preprocessing step is applied on the T_{plain} of 2-gram, 3-gram, 4-gram and 5-gram to filter out noisy data before computing W_i.

To generate the weight for an N-gram, it is treated as a single entity and fed to the weighted URL tokens system. If either N_c or N_a is a substring of T_{url}, the weighted URL tokens system will assign a weight to N_o based on Eq. (1). Note that N_c, N_a and T_{url} are lowercased prior to the string matching process.

After generating the weight for all N-grams, the N-gram with the highest weight is selected as the multi-gram identity keywords and denoted as Y. Note that it is also possible for all N-grams to have zero weight, which may indicate that the query webpage does not contain multi-gram identity keywords. Lastly, the target identity keywords are redefined as $K = X \cup Y$.

3.4. Search engine lookup

In this module, the target identity keywords will be fed into the search engine to find the target domain name. The rationale of using search engine is its ability in mapping a given set of identity keywords to its corresponding website URL. However, the search results may occasionally contain excessive information. To effectively single out the target domain name from the search results, several identity-relevant features are analysed using the compromise programming

Table 3
Some sample results of N-gram pre-processing.

N_o	Pre-processing results		T_{url}	Matched
	N_c	N_a		
Bank of America	bankofamerica	boa	bankofamerica.com	Yes
Internal revenue service	internalrevenueservice	irs	irs.gov	Yes
Hong Leong Bank	hongleongbank	hlb	hlb.com.my	Yes
Commonwealth Bank of Australia	commonwealthbankofaustralia	cboa	commbank.com.au	No
Sports illustrated	sportsillustrated	si	si.com	Yes
Entertainment weekly	entertainmentweekly	ew	ew.com	Yes

algorithm, which will be thoroughly discussed in Section 3.5.4. In this work, Google is selected as the search engine based on the prior study by [12] on anti-phishing.

To integrate Google search service into the proposed phishing detection system, the search terms are embedded into a search query URL, which is similar to what is seen on the address bar of the search result page when using a web browser. For instance, given three target identity keywords, represented by *A*, *B* and *C*. The search query URL will appear as https://www.google.com/search?q=A+B+C. The default search query URL returns only the top ten results. PhishWHO follows the recommendation by [25] to retrieve the top 30 search results. Hence, the search query URL is formatted as https://www.google.com/search?num=30&q=A+B+C.

To obtain the search results, a Hypertext Transfer Protocol (HTTP) request is sent to the search engine using the search query URL. The search engine will then return a response in the form of a HTML webpage containing the search results. The HTML webpage can now be parsed into a DOM tree to extract all URLs listed in the result entries.

3.5. Target domain name finder

In this work, a novel method is proposed to combine the values of three identity-relevant features in order to pinpoint the target domain name from the search results. With the target domain name, it is possible to establish the legitimacy of the query webpage by comparing it against the actual domain name. The details of each feature are discussed in this subsection.

3.5.1. Feature 1 — identity keyword density

Most webpages contain identity keywords that are scattered evenly across five different sections, namely the title, meta elements, URLs in HTML source code, visible text, and the page URL. On the other hand, non-identity keywords tend to have uneven distribution, in the sense that some can only be found in the visible text but not in other sections. Note that URLs in HTML source code include hyperlinks, paths of resources, and URLs in the action attributes of login forms. This distinctive pattern can be transformed into a useful feature to facilitate selection of the target domain name. Thus, the identity keyword density feature is proposed. In the feature generation process, the aforementioned five sections of a query webpage are analysed to find words that match the second level domains (SLDs) from the search results. SLD is a specific URL substring that precedes the top level domain (TLD), as shown in Fig. 5. Companies often use their brand names as the SLD string to help visitors in recognising their websites.

Table 4
Computation of identity keyword density.

Section index, *i*	Section	Can be found	Score
1	Title	Yes	$d_1 = 1$
2	Meta	No	$d_2 = 0$
3	SLDs	Yes	$d_3 = 1$
4	Visible text	Yes	$d_4 = 1$
5	Page URL	Yes	$d_5 = 1$
			Total score = 4
			$F_{density} = 0.8$

For every domain name in the search result, its identity keyword density $F_{density}$ is calculated as Eq. (2).

$$F_{density} = \frac{1}{n}\sum_{i=1}^{n} d_i, \tag{2}$$

where d_i denotes the existence of the identity keyword in section *i* of the query webpage (set to 1 if the identity keyword exists, otherwise 0), *i* represents the section index of the query webpage, and *n* is the total number of sections. An example of calculation for $F_{density}$ is shown in Table 4.

3.5.2. Feature 2 — frequency of domain name in search results

This feature is founded on the notion that the target domain name corresponding to the given target identity keywords is more likely to be listed more often in the search results when compared to other unrelated domain names. By denoting this feature value as F_{freq1}, we can observe in Table 5 that the website apple.com has the highest frequency ($F_{freq1} = 12$) among the search results when the terms "Apple ID" is searched on Google.

3.5.3. Feature 3 — frequency of domain name in query webpage

In this subsection, another identity-relevant feature is proposed, based on the same fact that the target domain name listed in the search results should have a higher frequency of occurrence in the HTML source of the query webpage. Hence, for each unique domain name in the search results, its number of occurrence in the query webpage HTML source code is counted and denoted as F_{freq2}. Based on the sample results in Table 5, apple.com shows a higher F_{freq2} compared to other unrelated domain names in the search result.

3.5.4. Compromise programming

Three identity-relevant features are proposed as the criteria to select the target domain name from the search results. All these features have different value ranges, thus combining their values through direct summation would introduce biased result. In addition, the importance of each feature must be taken into consideration. Theoretically, the ideal solution (i.e., target domain name) is

Table 5
Sample values of identity-relevant features from a real phishing webpage targeting Apple.

Domain name in search result	$F_{density}$	F_{freq1}	F_{freq2}
wikipedia.org	0.0	1	0
macworld.co.uk	0.0	3	0
lifehacker.com	0.0	1	0
apple.com	0.6	12	1.0
osxdaily.com	0.0	1	0
Identity keywords extracted = "Apple ID"			

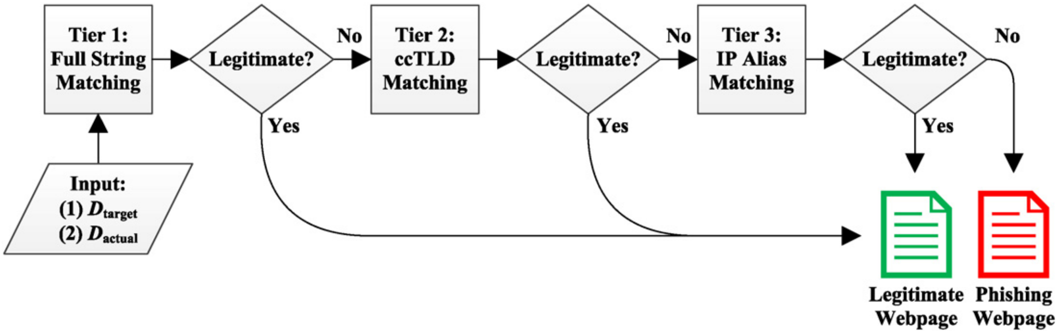

Fig. 4. Process flow of 3-tier identity matching system.

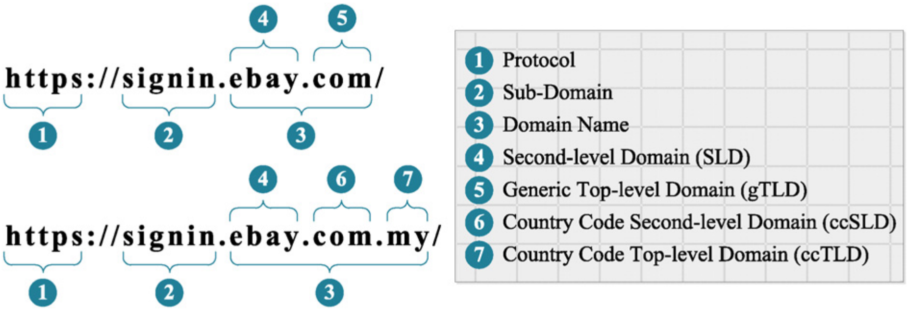
https://signin.ebay.com/
4
5
1
2
3
https://signin.ebay.com.my/
4
6
7
1
2
3
1 Protocol
2 Sub-Domain
3 Domain Name
4 Second-level Domain (SLD)
5 Generic Top-level Domain (gTLD)
6 Country Code Second-level Domain (ccSLD)
7 Country Code Top-level Domain (ccTLD)

Fig. 5. Anatomy of URL.

the domain name having the maximum F_{density}, F_{freq1} and F_{freq2}. In practice, however, the ideal solution may not exist, in cases where there is no domain name having the maximum value for all three features. This is where the well-established compromise programming method from the field of operations research can be applied to solve the problem of making an optimised decision in the presence of multiple features.

Compromise programming computes a value that reflects the overall distance to the ideal solution. Considering all criteria are to be maximised, the compromise programming distance metric in the context of target domain name selection is expressed as Eq. (3).

$$L_p(d) = \left[\sum_{j=1}^{J} w_j^p \left|\frac{f_j^* - f_j(d)}{M_j - m_j}\right|^p\right]^{\frac{1}{p}}, \qquad (3)$$

where d refers to the unique domain name, $L_p(d)$ is the distance metric for d, and J denotes the total number of features. The ideal value of feature j is denoted by f_j^*, while the value of feature j for d is represented by $f_j(d)$. The maximum (ideal) and minimum (inverse ideal) values of feature j among all d are denoted by M_j and m_j, respectively. The weight of feature j is denoted by w_j^p, while p is the parameter/balancing factor reflecting the attitude of the decision maker with respect to compensation between deviations. After computing all the required L_p values, the target domain name D_{target} is chosen based on the smallest L_p value.

3.6. 3-Tier identity matching

In this phase, the target domain name D_{target} and the actual domain name D_{actual} are passed into a 3-tier identity matching system. Each tier will produce a binary result indicating the phishing status of the query webpage. If the result is positive (i.e., phishing), the matching process continues to the subsequent tiers. Otherwise, the webpage is concluded as legitimate and the processing ends at the current tier. Fig. 4 shows the process flow of the 3-tier identity matching system.

One may argue that performing string comparison on D_{target} and D_{actual} alone is sufficient to conclude whether a webpage is a phishing webpage. In the proposed method, the analysis of webpage identity is taken two steps further. We claimed that there exist two sets of indirect identity relationships behind every domain name. These indirect identity relationships can be used to establish an identity match (i.e., legitimate result), even when D_{target} is not equal D_{actual} (related examples will be given in the following subsections). Therefore, a 3-tier identity matching system is proposed to discover these indirect identity relationships and conclude the legitimacy of the query webpage.

3.6.1. Tier-1 — full string matching

With D_{target} and D_{actual} in hand, full string matching is performed to find a match as shown in an example in Table 6. If D_{target} is identical to D_{actual}, the query webpage is concluded as legitimate.

Table 6
Tier-1 detection result.

Case	D_{target}	D_{actual}	Full string matched	Detection result
1	ebay.com.my	ebay.com.my	Yes	Legitimate
2	ebay.com.my	ebay.com	No	Phishing
3	ebay.com	ebay.co.uk	No	Phishing
4	ebay.com	auctionforamerica.org	No	Phishing
5	ebay.com	ebayturk.org	No	Phishing

Table 7
Tier-2 detection result.

Case	D_{target}	D_{actual}	SLD matched	ccTLD matched	Detection result
1	ebay.com.my	ebay.com.my	SKIP	SKIP	Legitimate
2	ebay.com.my	ebay.com	Yes	Yes	Legitimate
3	ebay.com	ebay.co.uk	Yes	Yes	Legitimate
4	ebay.com	auctionforamerica.org	No	SKIP	Phishing
5	ebay.com	ebayturk.org	No	SKIP	Phishing

Otherwise, the processing continues to the next tier of identity matching.

3.6.2. Tier-2 — ccTLD matching

Tier-2 identity matching exploits the indirect identity relationship between regular domain names and country-specific domain names. As shown in Fig. 5, the right-most segment in a domain name is referred to as the TLD. TLD can be classified into two main categories — generic TLDs (gTLDs) and country-code TLDs (ccTLDs). Examples of gTLDs include .com, .net, .org, and .biz.

On the other hand, ccTLDs consist of only two letters which represent the abbreviation of different countries or territories, such as .uk, .au, and .fr. The utilisation of ccTLDs is mainly for the country or language that the websites are based in. Hence, for a particular legitimate website, it might be accessible through different URLs such as https://www.amazon.com or https://www.amazon.co.uk. Both of these websites belong to Amazon. The only difference in the domain name is the ccTLD extension .uk for the Amazon United Kingdom website. Such application of ccTLD is very common nowadays, especially in multinational corporations. Hence, given a specific brand or organization, an indirect identity relationship is established by assuming its regular domain names and country-specific domain names to be the same identity and under a single ownership. Table 7 shows the improved detection result using tier-2 identity matching.

To verify that D_{target} or D_{actual} is indeed having a valid ccTLD, the tier-2 module performs a look-up to a ccTLD list. This ccTLD list is published in the Root Zone Database[1] of the Internet Assigned Numbers Authority (IANA). Based on the tier-2 detection result, case-2 and case-3 can now be correctly labelled as legitimate. However, the correct detection result for case-4 should be legitimate instead of phishing. This is due to the fact that auctionforamerica.org and ebay.com are actually under the ownership of eBay Inc. To account for such cases of identity discrepancy, the processing continues to the final tier of identity matching.

3.6.3. Tier-3 — IP alias matching

The final phase of the identity matching system is based on IP alias matching. IP alias matching exploits the relationship between domain names and IP addresses to match D_{target} and D_{actual}. Fig. 6 shows the existence of many-to-one relationship between domain names and IP addresses.

Many-to-one relationships are widely used in the web architecture, allowing different websites belonging to an organization's subsidiaries to be hosted on a single web server. This relationship indirectly associates different websites under the same ownership. Based on this relationship, the domain name D_1, D_3 and D_4 can be considered as matching since they point to the same IP address or web server. Applying IP alias matching on case-4 produces a correct detection result as shown in Table 8. As for case-5, no indirect identity relationship can be established by the end of the final tier. Therefore, the tier-3 module concludes case-5 as phishing and completes the processing.

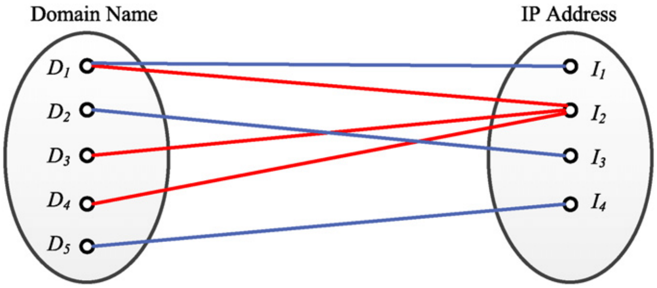

Fig. 6. Existence of many-to-one relationship between domain names and IP addresses.

It is important to note that tier-2 and tier-3 modules are interdependent. For example, it is not possible to match hsbc.com.my and hsbc.co.uk using tier-3 module alone, since both domain names do not possess any matching IP alias. However, it is possible to yield a match using tier-2 module. Hence, the absence of tier-2 module will certainly increase the false positive rates.

4. Results and analysis

In this section, the experimental results of the proposed phishing detection method are presented and compared against the results achieved by the conventional methods. Specifically, 5000 phishing webpages based on URLs from PhishTank[2] and OpenPhish[3] , and another 5000 legitimate webpages based on URLs from Alexa[4] , i.e., top one million list during the period from January to May 2015, are selected for experiment purpose. To make local copies of the webpages, the GNU Wget[5] tool driven by an automated Python script is used. The complete HTML documents are downloaded together with the resources (e.g., images, CSS, JavaScript) for proper webpage display purpose. In addition, the screenshot of every downloaded webpage is saved. This is necessary to accelerate the manual inspection and filtering process without having to open up each local copy of the webpage in a browser.

The webpages were further processed. First, there are instances of webpage which failed to load in both the phishing and legitimate dataset. Second, there are also some instances of webpage in our phishing dataset that are actually legitimate. The URL of these legitimate websites ended up in the phishing feeds of PhishTank or OpenPhish when phishers hijacked them and injected phishing webpages into these websites. Subsequently, these compromised websites have recovered to serve its original legitimate content by the time of download. Third, some phishing webpages hosted on free hosting providers have been taken down, thus accessing its URL leads to "error 404 page" or the default landing page of the free hosting providers. These types of invalid webpages in the downloaded

[1] http://www.iana.org/domains/root/db/

[2] http://www.phishtank.com/
[3] https://www.openphish.com/
[4] http://www.alexa.com/
[5] http://www.gnu.org/software/wget/

Table 8
Tier-3 detection result.

Case	D_{target}/IP alias	D_{actual}/IP alias	IP alias matched	Detection result
1	ebay.com.my	ebay.com.my	SKIP	Legitimate
2	ebay.com.my	ebay.com	SKIP	Legitimate
3	ebay.com	ebay.co.uk	SKIP	Legitimate
4	ebay.com 66.211.160.86 66.211.160.87 66.135.216.190	auctionforamerica.org 66.135.216.190 66.211.160.88 66.211.160.87 66.135.210.101	Yes (66.211.160.87)	Legitimate
5	ebay.com 66.211.160.86 66.211.160.87 66.135.216.190	ebayturk.org 74.220.199.8	No	Phishing

Table 9
Evaluation results for Experiment I.

Combination	TPR (%)	FPR (%)	TNR (%)	FNR (%)	Accuracy (%)	MCC value
Tier-1	99.68	10.90	89.10	0.32	94.39	0.8928
Tier-1 + Tier-2	99.68	7.74	92.26	0.32	95.97	0.9219
Tier-1 + Tier-2 + Tier-3	99.68	7.48	92.52	0.32	96.10	0.9244

dataset can adversely affect the integrity of the evaluation results. Hence, a thorough filtering process is conducted to remove them. Duplicate instances of webpages are also removed.

All experiments were conducted on a desktop computer equipped with an Intel Pentium D 2.8 GHz CPU, 1 GB RAM and Windows 7 Professional 32-bit operating system.

4.1. Experiment I — evaluation of 3-tier identity matching system

Experiment I is designed to evaluate the performance of PhishWHO using three combinations of the modules in the 3-tier identity matching system as shown in Table 9. The true positive rate, true negative rate, false positive rate, and false negative rate are denoted as TPR, TNR, FPR and FNR, respectively. Results suggest that PhishWHO achieves a stable TPR of 99.68% in all combinations, while the FPR reduces progressively as tier-2 and tier-3 modules are being added.

To further compare the overall performance, the Matthew's Correlation Coefficient (MCC) is considered. MCC is a metric for assessing the quality of the predicted value to the actual value, and its values range from −1 to +1 [4]. When MCC approach unity, it is indicative that the system achieves near to perfect prediction. Eq. (4) shows the calculation of MCC.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \tag{4}$$

where true positives, true negatives, false positives, and false negatives are denoted as TP, TN, FP and FN, respectively. Results suggest that the combination of tier-1, tier-2 and tier-3 modules is the best, achieving the highest MCC value of 0.9244.

4.2. Experiment II — conventional methods with similar analysis techniques

In Experiment II, the proposed method as well as two selected conventional methods with similar analysis techniques, i.e., text-based method [25] and identity-based method [20] were implemented. To facilitate the presentation, these methods are referred to as M1 and M2, respectively.

Results for Experiment II are recorded in Table 10. It is observed that PhishWHO outperforms the conventional methods in detecting phishing webpages, achieving a TPR of 99.68%. M2 is ranked second at a TPR of 96.22%. The least effective method in detecting phishing webpages is M1, which only scored a TPR of 88.82%.

As for legitimate webpage detections, M2 scored the highest TNR at 92.62%, while PhishWHO is slightly inferior with a TNR of 92.52%. M1 is once again outranked by all its competitors, showing the lowest TNR (i.e., 91.74%). However, the MCC values suggest that PhishWHO achieves better overall performance, followed by M2 and M1.

4.3. Experiment III — feature-based method

In Experiment III, the proposed method is benchmarked against a conventional feature-based method, i.e., [12], denoted as M3. Results are depicted in Table 11. It is evident that PhishWHO outperforms M3 in terms of TPR and achieves comparable TNR. Analysis using MCC values also proves that PhishWHO is more advantageous than M3.

5. Discussion and limitation

PhishWHO leverages the novel weighted URL tokens system based on N-gram model, which offers the advantage of classifying

Table 10
Benchmark results for Experiment II.

Phishing detection method	TPR (%)	FPR (%)	TNR (%)	FNR (%)	Accuracy (%)	MCC value
PhishWHO	99.68	7.48	92.52	0.32	96.10	0.9244
M1	88.82	8.26	91.74	11.18	90.28	0.8059
M2	96.22	7.38	92.62	3.78	94.42	0.8890

Table 11
Benchmark results for Experiment III.

Phishing Detection Method	TPR (%)	FPR (%)	TNR (%)	FNR (%)	Accuracy (%)	MCC value
PhishWHO	99.68	7.48	92.52	0.32	96.10	0.9244
M3	95.72	3.76	96.24	4.28	95.98	0.9196

webpages in different languages written in ASCII characters. Unlike M1 and M2 that depend on English-specific language semantics to extract identity keywords, PhishWHO depends on the proposed weighted URL tokens system to extract identity keywords. Therefore, PhishWHO is able to overcome most of the language limitation issues faced by M1 and M2.

5.1. *Effectiveness on existing phishing strategies*

This section discusses the impact of various existing phishing strategies to the effectiveness of PhishWHO. First, when phishers clone a website using phishing toolkit, the content (e.g., keywords, meta data) of phishing webpages will be exactly the same as the original legitimate webpages. Based on experimental results, most phishing samples which are exact clones of the legitimate website are successfully detected by the proposed method. This is because legitimate webpages tend to have more identity properties, thus it is in fact more favourable to PhishWHO if a phishing webpage appears more similar to the legitimate webpage.

From the experiment dataset, it is observed that the amount of phishing webpages hosted on hijacked legitimate websites is non-trivial. M1 recorded the lowest TPR among the four methods that are benchmarked. This is due to the fact that M1 often misclassified phishing webpages hosted on hijacked legitimate websites. M3 is also susceptible to the abuse of search engine ranking metrics by hijacked legitimate websites, since it gathers two features from search results (i.e., webpage ranking and total number of results). PhishWHO, on the other hand, is resistant against this phishing strategy due to the usage of three identity-relevant features instead of depending solely on search engine ranking metrics.

Phishers may also use visual cloning strategy to create webpages that consist of mostly images. In such cases, PhishWHO may extract insufficient identity keywords to be queried on search engine. For example, some phishing webpages are known to exploit images to replace the textual content of the whole webpage. This limitation can be resolved by invoking an Optical Character Recognition (OCR) module to extract text in the images and feed them to PhishWHO.

If the adversaries perform pharming or poison the DNS servers, the proposed solution may not be able to detect the authenticity of phishing websites. Pharming exploits vulnerabilities in Internet infrastructure to allow the phishing webpage to use domain name of the legitimate website (this is done at the DNS resolution level). Hence, the proposed solution will yield a match (i.e., legitimate outcome) when comparing the domain name of the phishing webpage to the targeted domain name. In our future work, this vulnerability can be addressed by adding forward and reverse DNS queries to validate the domain name and IP address of the query webpage. Another solution for pharming attacks is by leveraging multiple DNS servers to establish redundancy, assuming that the selected DNS servers are not compromised simultaneously.

Lastly, some phishing websites may install malware (e.g., Trojan) on the user's computer once the browser loads the phishing webpage. Thus, deploying the proposed method as a browser-based phishing detection system may not be sufficient. A better approach will be to incorporate the proposed method as a component into the anti-virus solution where the protection mechanism is more comprehensive and able to access the quarantine feature while blocking any incoming download from the phishing website.

6. Conclusion

In this work, a phishing webpage detection technique called PhishWHO is proposed, where the differences between the target and actual identities of a webpage are exploited for classification. Identity keywords are extracted from the textual contents of the webpage using the proposed novel weighted URL tokens system based on the N-gram model. These keywords are then searched using a search engine, and the returned domain name are compared against the suspicious domain name using the proposed 3-tier identity matching system, including full string matching, ccTLD matching and IP alias matching. Results suggest that PhishWHO outperforms the conventional phishing detection methods considered. The practical implications of this research include: (a) promoting secure and safe electronic communication channels to the society; (b) enhancing the confidence of consumers in using e-commerce services, and; (c) minimising financial losses faced by consumers and businesses as a result of phishing.

In future, it is worthwhile to consider research areas such as classification of webpages containing non-ASCII characters (e.g., Chinese, Japanese, Russian). In addition, we may enhance our proposed method with an OCR module to address visual cloning problems.

Acknowledgements

The funding for this project is made possible through the research grant obtained from UNIMAS and the Ministry of Education, Malaysia under the Fundamental Research Grant Scheme 2/2013 [Grant No: FRGS/ICT07(01)/1057/2013(03)]. This work was also partly supported by the University of Malaya Research Grant FL011-2011 (account number 5659961) under the purview of ICT & Computational Science Research Cluster, UM Research.

References

[1] M. Alsharnouby, F. Alaca, S. Chiasson, Why phishing still works: user strategies for combating phishing attacks, International Journal of Human-Computer Studies 82 (2015) 69–82.
[2] APWG, Phishing Activity Trends Report, 2nd Quarter 2014, 2014, retrieved from http://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf. Accessed 26.03.16.
[3] APWG, Global Phishing Survey: Trends and Domain Name Use in 2H2014, 2015, retrieved from http://docs.apwg.org/reports/APWG_Global_Phishing_Report_2H_2014.pdf. Accessed 26.03.16.
[4] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (5) (2000) 412–424.
[5] I. Bose, A.C.M. Leung, Unveiling the mask of phishing: threats, preventive measures, and responsibilities, Communications of the Association for Information Systems 19 (1) (2007) 24.
[6] I. Bose, A.l.C.M.a.n. Leung, Do phishing alerts impact global corporations? A firm value analysis, Decision Support Systems 64 (2014) 67–78.
[7] K.L. Chiew, E.H. Chang, S.N. Sze, W.K. Tiong, Utilisation of website logo for phishing detection, Computers & Security 54 (2015) 16–26.
[8] R. Dhamija, J.D. Tygar, M. Hearst, Why phishing works, Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems, Montréal, Québec, Canada, ACM. 2006, pp. 581–590.

[9] DigiCert Inc, DigiCert Phishing White Paper: A Primer on What Phishing is and How It Works, 2009, retrieved from http://www.digicert.com/news/DigiCert_Phishing_White_Paper.pdf. Accessed 26.03.16.

[10] E.M.C. Corporation, RSA Monthly Fraud Report, 2015, retrieved from http://australia.emc.com/collateral/fraud-report/h13929-rsa-fraud-report-jan-2015.pdf. Accessed 26.03.16.

[11] A.Y. Fu, L. Wenyin, X. Deng, Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD), IEEE Transactions on Dependable and Secure Computing 3 (4) (2006) 301–311.

[12] J.H. Huh, H. Kim, Phishing detection with popular search engines: simple and effective, 4th Canada-France MITACS Workshop on Foundations and Practice of Security, Paris, France, May 12-13, 2011, Springer Berlin Heidelberg. 2011, pp. 194–207. ISBN 978-3-642-27901-0.

[13] K. Komiyama, T. Seko, Y. Ichinose, K. Kato, K. Kawano, H. Yoshiura, In-depth evaluation of content-based phishing detection to clarify its strengths and limitations, Proceedings of the International Conference on U- and E-Service, Science and Technology, Jeju Island, Korea, December 13–15, 2010, Springer Berlin Heidelberg. 2010, pp. 95–106. ISBN 978-3-642-17643-2.

[14] A. Le, A. Markopoulou, M. Faloutsos, PhishDef: URL names say it all, Proceedings of IEEE INFOCOM, Shanghai, China, April 10-15, 2011, IEEE. 2011, pp. 191–195.

[15] W. Liu, N. Fang, X. Quan, B. Qiu, G. Liu, Discovering phishing target based on Semantic Link Network, Future Generation Computer Systems 26 (3) (2010) 381–388.

[16] W. Liu, G. Liu, B. Qiu, X. Quan, Antiphishing through phishing target discovery, IEEE Internet Computing 16 (2) (2012) 52–61.

[17] MarkMonitor Inc, Phishing trends: who are fraudsters targeting now? 2015, retrieved from https://www.markmonitor.com/mmblog/phishing-trends-who-are-fraudsters-targeting-now/. Accessed 26.03.16.

[18] R.M. Mohammad, F. Thabtah, L. McCluskey, Tutorial and critical analysis of phishing websites methods, Computer Science Review 17 (2015) 1–24.

[19] S. Purkait, Examining the effectiveness of phishing filters against DNS based phishing attacks, Information and Computer Security 23 (3) (2015) 333–346.

[20] G. Ramesh, I. Krishnamurthi, K.S.S. Kumar, An efficacious method for detecting phishing webpages through target domain identification, Decision Support Systems 61 (2014) 12–22.

[21] S. Sheng, B. Wardman, G. Warner, L.F. Cranor, J. Hong, C. Zhang, An empirical analysis of phishing blacklists, Proceedings of the Conference on Email and Anti-Spam (CEAS), Mountain View, California, USA, July 16-17, 2009, 2009.

[22] C.L. Tan, K.L. Chiew, S.N. Sze, Phishing website detection using URL-assisted brand name weighting system, Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Kuching, Malaysia, December 1–4, 2014, IEEE. 2014, pp. 54–59.

[23] H. Zhang, G. Liu, T.W.S. Chow, W. Liu, Textual and visual content-based anti-phishing: a Bayesian approach, IEEE Transactions on Neural Networks 22 (10) (2011) 1532–1546.

[24] Y. Zhang, S. Egelman, L. Cranor, J. Hong, Phinding phish: evaluating anti-phishing tools, Proceedings of the 14th Annual Network & Distributed System Security Symposium, San Diego, California, USA, February 28–March 2, 2007, ISOC. 2007.

[25] Y. Zhang, J.I. Hong, L.F. Cranor, Cantina: a content-based approach to detecting phishing web sites, Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, May 8–12, 2007, ACM. 2007, pp. 639–648.

Choon Lin Tan received his BEng degree in Electronics and Computer Engineering from Universiti Malaysia Sarawak (UNIMAS). He is currently pursuing his MSc degree in the Faculty of Computer Science and Information Technology. His research interests include web document analysis, information retrieval and anti-phishing.

Kang Leng Chiew is currently a senior lecturer at the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS). He received his PhD in Computer Science specialised in Information Hiding from Macquarie University, Sydney, Australia. His research interest is in information security. He is currently working in anti-phishing research. Past researches include steganalysis on digital images. Previously, he also worked in the area of image processing.

KokSheik Wong received the dual B.Sc. and M.S. degrees in computer science and mathematics from Utah State University, Logan, UT, USA, in 2002 and 2005, respectively, and the D.Eng. degree from Shinshu University, Matsumoto, Japan, in 2009, under the scholarship of Monbukagakusho. In 2010, he joined the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia, where he is currently a Senior Lecturer. He is a member of the Centre for Image and Signal Processing with the University of Malaya, where he leads the Multimedia Signal Processing and Information Hiding Group. His research interests include information hiding, steganography, watermarking, multimedia perceptual encryption, multimedia signal processing, and their applications.

San Nah Sze received her BSc and MSc from University Technology Malaysia, and PhD from Sydney University. She is now a senior lecturer at University Malaysia Sarawak. Her research interests include vehicle routing, manpower planning, shift scheduling, heuristic solution and timetabling.