SECURITY AND COMMUNICATION NETWORKS
Security Comm. Networks 2015; **8:**888–898
Published online 8 August 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/sec.1045

RESEARCH ARTICLE

Combating phishing attacks via brand identity and authorization features

Guang-Gang Geng[1]*, Xiao-Dong Lee[1] and Yan-Ming Zhang[2]

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100180, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, 00180, China

ABSTRACT

Phishing, also called brand spoofing, has become the most troubling scam on the Internet, which seriously threatens the Web security. The essence of phish is that "robbers" use false sites, which look like a trustworthy brand site, where favicon, logo and copyright notice are important brand identities. We analyzed 78-day phishing data of PhishTank and Anti-Phishing Working Group (APWG). The statistics show that more than 98.93% phishing sites contain at least one brand entity—favicon, logo or copyright notice. Indeed, only a few lowest-quality phishing campaigns do not use such brand elements. Obviously, brand entities are powerful weapons of phishers to trick users. By analyzing the characteristics of brand entities in phishing sites, several brand identity features are extracted. However, only brand entities do not consider whether the Web page with brand entities belongs to the corresponding brand or has an authorization to use the brand entities. To solve this problem, redirection, incoming links and Domain Name System (DNS) information-based brand authorization features are further extracted to discriminate the sites with branding rights from phishing sites. Based on extracted features, statistical anti-phishing classification models are trained. We collected a diverse spectrum of corpora containing 3863 phishing cases from PhishTank and APWG, and 17 571 legitimate samples from DMOZ, Google and DNS resolution log. Experimental evaluations show that the model achieves 98.8% true positive rate and 0.09% false positive rate, which demonstrates the competitive performances of extracted features for statistical anti-phishing in practice. Copyright © 2014 John Wiley & Sons, Ltd.

KEYWORDS

brand identity recognition; phishing attacks; brand authorization feature; machine learning

***Correspondence**

Guang-Gang Geng, Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100180, China.
E-mail: gengguanggang@cnnic.cn

1. INTRODUCTION

Phishing is a scam typically carried out by unsolicited email and/or Web sites that pose as legitimate sites and lure unsuspecting victims to provide personal and financial information. Nowadays, phishing has spread beyond email to include instant messaging, voice over Internet protocol (IP), false advertising, social media and even massively multiplayer online games [1]. In this paper, we use a general definition of phishing given by Whittaker *et al.* [2]. They defined a phishing page as any Web page that, without permission, alleges to act on behalf of a brand with the intention of confusing viewers into performing an action with which the viewer would only trust a true agent of the brand.

Although there has been an increase in the general public awareness of online security, phishing is still a major threat to the netizen. Dhamija *et al.* claimed that high-quality phishing sites could fool 90% of users [3]. Phishing attacks can lead to damaging losses in terms of identity theft, sensitive intellectual property, corporate secrets and national-security secrets [1]. The direct losses that caused by phishing are more than $1bn per year in the USA [1,3].

Given the risks of phishing attacks, many academic research and business practices have been performed. Among them, the most popular phishing countermeasures include manually verified blacklists, and heuristic learning-based or machine learning-based methods. The blacklist-based method is best known for the Web browsers such as Internet Explorer, Mozilla Firefox, Safari and Opera, which achieves fairly low false positive rate (FPR), but is ineffective for fresh phishes [1,2,4]. For learning-based methods, multiple features such as URL [2,5,6], content (title text, hyperlink, logo, form, etc.) in Web

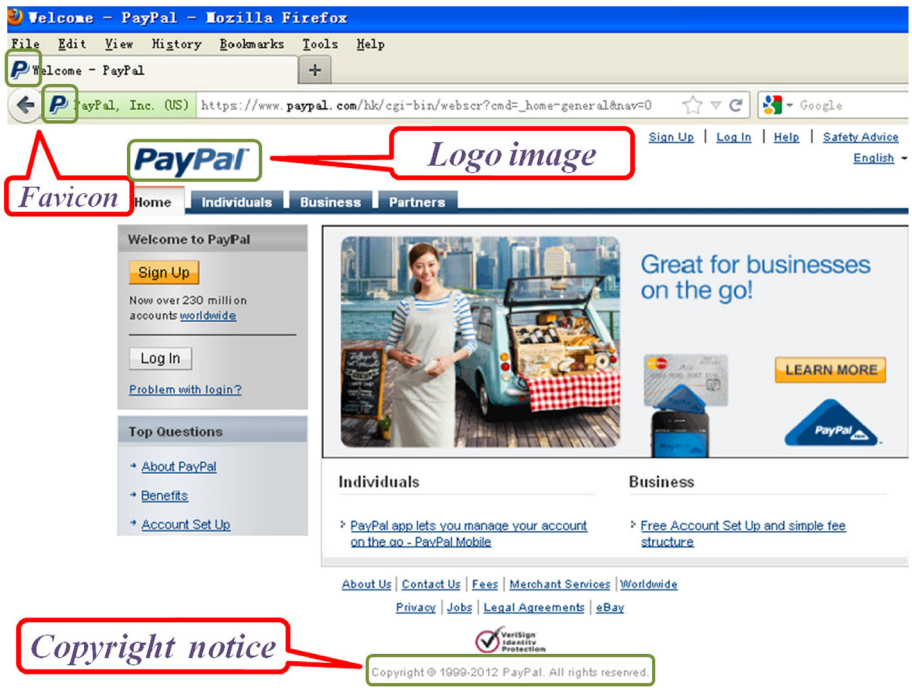

Figure 1. Paypal's favicon, logo and copyright notice shown in Firefox 25.0.

pages [2,5,7] and third-party services (PageRank, search engines, WHOIS, etc.) [5,8] are used to learn the phishes detection models.

This paper proposes a statistical phishing detection method with brand features. The premise behind the study is that almost all phishing sites are designed to look like a trustworthy third party to fool the users, where brand spoofing is the most important features. Figure 1 shows the favicon, logo and copyright notice of *paypal.com*.

We analyzed 78-day phishing data of PhishTank [9] and Anti-Phishing Working Group (APWG) [10]. The statistics show that 80.7% phishing sites employ fake favicons, 86.2% phishing sites contain brand logos and 84.1% phishing sites have copyright notice of corresponding brands. And surprisingly, more than 98.93% phishing sites contain at least one brand entity—favicon, logo or copyright notice. Indeed, only a few lowest-quality phishing campaigns do not use such brand elements. In other words, brand entities are powerful weapons of phishers to trick users. However, existing statistical learning-based anti-phishing research has not yet paid attention to the fact.

The favicon is displayed on major Web browsers' favorites menu, address bar, bookmarks and page tabs. In fact, favicon is rapidly becoming an important element of brand identity online. More and more Internet users treat it as the symbol of a company. The criminals become aware of the importance of favicons. Our statistics show that most of phishing sites provide identical or similar favicons of their target brands to mislead users. Taking the two most-attacked targets, *paypal.com* and *taobao.com*, as an example, more than 99% of their pretenders take advantage of favicons.

A logo is a graphic mark or emblem, which is commonly used by commercial enterprises, organizations and even individuals to aid and promote instant public recognition. Because of the significance of logos, almost all the company's Web sites contain a logo. Our statistics show that quite a few phishing sites provide identical or similar logos of their target brands to deceive users.

A copyright notice informs users of the underlying claim to copyright ownership in a published work. The vast majority of Web sites have a copyright notice in the footer. The notice for visually perceptible copies should contain three elements: the symbol © (the letter c in a circle), or the word copyright; the year of first publication; and the name of the owner of copyright. Most designers do this as routine on all Web sites they design, and most of users treat the copyright notice as a brand identifier of a reputable Web site. A number of phishing sites include copyright notice in the footer.

In this paper, by analyzing the characteristics of the visual brand entities—favicon, logo and copyright notice—in phishing Web sites, a series of brand features are extracted, and furthermore, brand authorization features are extracted to discriminate legitimate brand sites or sites with branding rights from spoofing ones.

Our main contributions are summarized as follows:

- This paper proposes a series of brand identity features to combating phishing attacks. To the best of our knowledge, this is the first paper that presents a

unified analysis of favicon, logo and copyright notice in phishing sites.
- Several brand authorization features are extracted to discriminate the legitimate brand sites from phishes, which reduces the FPR of favicon-based anti-phishing method.
- The proposed features are language independent.† In other words, no matter what internationalized domain and what language a phishing site uses, as long as the site contains the brand entities, it is in the range of detection.
- Experimental evaluations on a complex data set demonstrate the effectiveness of the proposed brand identity and authorization features-based statistical phishing detection method.

The rest of sections are organized as follows. Section 2 presents a literature review. Section 3 describes the brand entities detection and features representation. In section 4, we elaborate the brand authorization features to optimize the detection results. Section 5 describes the data sets and gives detailed evaluation. Lastly, section 6 draws the conclusion and provides some implications for the future work on anti-phishing.

2. RELATED WORK

In previous years, email is the main spread way of phishing. Most of the anti-phishing researches are focused on phishing email detection [11–13]. Filtering phishing email is regarded as the first line of defense of preventing phishing attacks from reaching end users [3]. However, nowadays, phishing attacks are increasingly sophisticated. Beyond email, phishing attacks spread via twitter direct messages [14], games on social networks [15], voice over IP phones [16] and more; thus, blocking the phishing attacks from the source becomes increasingly difficult.

In recent years, more research work on anti-phishing has been performed [1–8,17–20]. The ways of anti-phishing include browser-based anti-phishing tools, using Secure/Multipurpose Internet Mail Extensions (S/MIME), biometrics authorization, heuristic rules-based anti-phishing, statistical learning-based phishing detection and so on. In this section, we will focus on statistical learning-based methods.

The core of statistical learning-based anti-phishing methods is to recognize the potential patterns of phishing sites in different features. The extracted features include URL, visual information, page content and third-party services [2,5–8].

To capture the patterns in phishing URLs, several URL-based anti-phishing methods are proposed [6,21]. They all claimed achieving an accuracy of over 90%. However, URLs could be manipulated with little cost, which leads to the light-weighted method having unstable performance [5].

Xiang *et al.* proposed a hybrid phishing detection method based on identity recognition and third-party information retrieval [18]. The method requires no training data, no prior knowledge of phishing signatures and specific implementations and thus is able to adapt quickly to constantly appearing new phishing patterns. Though interesting, their experimental results with a true positive rate (TPR) of 90.06% with an FPR of 1.95% are not convincing, which is limited by the progress of information extraction technique research.

A recent feature-based work CANTINA+ used a rich set of features, including URL, HTML structure and some third-party services features [5]. The paper depends too heavily on third-party services, which reduces the overall detection efficiency. On the other hand, the corpus only includes English Web pages for their content-based features, and the proportion of legitimate samples in the corpora is too small. In reality, ratio of phishing and legitimate sites is imbalance. The 92% TPR on unique testing set is unconvincing for a binary classification.

To exploit visual similarity, Dhamija *et al.* proposed a method named dynamic skins, which focuses on image identity verification [19]. However, the paper does not give a formal evaluation. Maurer *et al.* presented a framework that uses visual Web site similarity to detect possible phishing Web sites and to create better warnings for such attacks [22]. They did not give a experimental evaluation either. Wang *et al.* proposed a logo recognition-based phishing detection method [20]. Its premise is that a critical element in virtually all fraudulent sites is a brand logo of the institution being imitated. However, merely employing the logo is not robust, for example, the pages containing login portals of the world's most-attacked target—*taobao.com* [23]. Figure 2 shows a screenshot of a typical phishing page targeting at *taobao.com*. Besides, the paper does not give an effective filter method for navigation Web sites and news Web sites, which often contain some brand logos.

In our previous work [24], we analyzed the popularity of favicons in phishing sites, proposed a heuristic rule-based favicon recognition algorithm and found that favicon is a good clue to detect phishing sites with spoofing favicons. However, not all the phishing sites have a fake favicon. Besides, detecting phishing only via favicon is easy to be bypassed. In this paper, instead of heuristic Boolean estimation, we first evaluate the favicon similarity to a confidence interval and then train a robust model by a series of brand identity and authorization features, where favicon is one of the important features.

Brand identities spoofing has become a powerful weapon of phishers. However, existing anti-phishing research pays little attention to them. This paper focuses on the widely used brand elements—favicon, logo and

† In this paper, all the proposed features are language independent, except for copyright notice feature.

DOI: 10.1002/sec

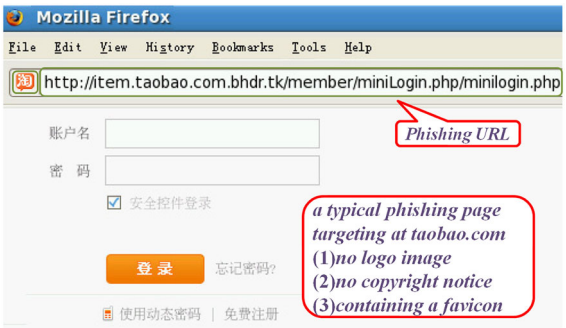

Figure 2. A typical phishing page targeting at taobao.com, shown in Firefox 24.1.

Figure 3. Favicons of the most targeted phishing brands.

copyright notice. A series of brand identity and authorization features are extracted, which are then used to train a robust statistical anti-phishing model.

3. BRAND ENTITIES DETECTION AND FEATURES REPRESENTATION

Favicon, logo and copyright notice are the most important brand identities of company sites, which are widely used by phishers to trick the users. In this section, we first analyze the characteristic of favicon, logo and copyright notice in phishing sites and then discuss the brand identity features representation.

3.1. Favicon feature extraction

Figure 3 shows favicons of the most targeted phishing brands.‡ The most targeted industries include financial banks, online payment services, insurance companies, governments, email services, hotels, security services, social network sites, retail services and auctions.

Table I illustrates the ways that a favicon can be recognized by the Web browsers.

‡ All the favicons were collected in October 2011. Although favicons change very infrequently, they are not set in stone for the long term.

Table I. Different ways of associating the favicon with a Web page.

<link rel="shortcut icon" href="http://example.com/image.ico" / >
<link rel="icon" type="image/vnd.microsoft.icon" href="http://example.com/image.ico" />
<link rel="icon" type="image/png" href="http://example.com/image.png" />
<link rel="icon" type="image/gif" href="http://example.com/image.gif" />
The favicon file named *favicon.ico* is located in the Web site's root directory.

Based on Table I, favicon files can be detected. The Web page source code should be parsed first, which is based on the fact that the Web browsers prefer (X)HTML link tag-specified image to root directory [25].

In general, the favicon is an ico file, which can contain one or more small images, each with a different size and/or color depth. Besides ico file, the JPEG, pnd, gif, apng and svg are also allowable favicon file formats.

To compare a suspicious favicon with a brand favicon, the first thing to do is extracting all the images and then resizing all the images to 16×16 pixels. The motivation of resizing all favicon images to 16×16 pixels is that 256 pixels is the most popular favicon size, and the icons shown on the favorites menu, the address bar and a page tab are all 16×16 pixels visual area. The image scaling is not limited to any resizing algorithms. In our experiments, nearest-neighbor interpolation algorithm is used [26]. Taking into account that 256 pixels images are relatively small, any difference of such two images will be amplified. In this paper, we measure the similarity of two favicon images by matching their histograms. In our experiments, we use correlation metric to compute the histogram similarity.

For a given suspicious URL, if the favicon exists, $F_{favicon}$ is computed as follows; otherwise, the favicon feature $F_{favicon} = 0$.

$$F_{favicon} = \max_i d\left(fav, fav^i\right) = \max_{i,j,k} sim\left(img^j, img_k^i\right) = \frac{\Sigma_I \left(H^j(I) - \overline{H^j}\right)\left(H_k^i(I) - \overline{H_k^i}\right)}{\sqrt{\Sigma_I \left(H^j(I) - \overline{H^j}\right)^2 \left(H_k^i(I) - \overline{H_k^i}\right)^2}} \quad (1)$$

where fav^i is an element of brands' favicon set, img^j is an image in favicon fav and img_k^i is an image in favicon fav^i. H^j and H_k^i are the corresponding gray histograms of img^j and img_k^i.

3.2. Logo extraction and feature representation

Unlike favicon, no agreement or comment dictates where to place a logo. That is, a logo can be displayed anywhere, which makes locating a logo image very difficult. Wang *et al.* extracted all the images embedded with <img> tags in HTML [20]. Although it is easy to realize, the method is time consuming, especially when the page contains many images. And worse still, many news and portal sites often contain many brand logos. In this case, a high FPR is unavoidable.

One natural question arises: how can we accurately locate the logos? Before answering that, let us see a common pattern noticed by Eyetrack III researchers: The eyes most often fixated first in the upper left of the page then hovered in that area before going left to right [27]. When significant content is outside that key upper left corner, it may be virtually invisible when people are making the big decision: whether to read more or quit the page. The phishers also know the pattern well. They usually put the most important identity—logos on the left top of the pages to trick the users. Based on the same pattern, we use vision-based page segmentation algorithm to locate the images in the top left of the page as logo candidates [28].

For a given URL, the page may contain more than one image in the top left. That is, multiple logo candidates can be detected. So the first thing is to delete the obvious noisy images, such as 1×1, 1×2 and 2×2 images, which are often used in Web design.

In the logo extraction step, besides the logo candidate itself, the URL of the image is also extracted. The reason for this is that statistics show that many phishers directly employ the logo image URL of the phished brand sites. Once the extracted URL is matched with one brand logo URL, the image matching step can be omitted, which will speed up the phishing detection.

In the rest of this section, we focus on logo matching. Just as favicon, logo files also have multiple formats, such as PNG, JPG and JIF. The ideal situation of "One Brand One Logo" does not also exist. Perhaps, the most important feature of the logo is that almost all famous brands have more than one logo. Figure 4 presents several famous brands of phishing that own multiple favicons.

The logo images in Figure 4 are all legitimate brand logos that used on different occasions, or at one time and another. They have obvious visual difference but have the same semantics. To carry out logo detection and recognition, the first thing to do is making a collection of phishing-targeted logos—*logoSet*. To minimize the problem of "Semantic Gap" in logo recognition, the basic principle of logo collection is attempting to cover all the common logo/logo-like images of the phished brands. The URLs of logo images in *logoSet* are also collected to *logoUrlSet*, if they exist.

In section 3.1, the gray histograms are used to matching favicons, which is based on the fact that favicon images are small enough that any flex or rotation is obvious for eyes.

Figure 4. Multiple logos of several famous phishing brands.

Figure 5. Screenshot of a typical phishing page footer targeting at Lloyds TSB Bank plc.

Can this method be used to recognize logos? The answer is no. Afroz *et al.* claimed that only 54% of phishing sites are detected via comparing binary equivalence and stated that "When logo-detection fails, it is because some logos are resized or the design is slightly changed in a way that is unnoticeable to the naked eye" [29]. In this paper, we compare logos via Hu moments [30], which are proven to be invariant to the image scale, rotation and reflection, except the seventh one, whose sign is changed by reflection. Hu moment invariants are widely used in pattern recognition [31]. We compare image A and image B via Hu moments as follows:

$$I(A,B) = \sum_{i=1\cdots7} \frac{\left|m_i^A - m_i^B\right|}{m_i^A} \tag{2}$$

where $m_i^A = sign(h_i^A) \times \log h_i^A$, $m_i^B = sign(h_i^B) \times \log h_i^B$, and h_i^A and h_i^B are the Hu moments of A and B, respectively.

For a given suspicious URL, we first parse the content of *URL* via vision-based page segmentation algorithm [28] to extract the images in the top left of the page and area bigger than 256 to *detLogoSet* and to extract the corresponding URLs to *detLogoUrlSet*. If *detLogoSet* is null, $F_{logo} = max$. If the intersection of *detLogoUrlSet* and *logoUrlSet* is not empty, $F_{logo} = 0$. In other cases, F_{logo} is computed as follows:

$$F_{logo} = \min_{i,j} I(logo_i, brandlogo_j) \tag{3}$$

where $logo_i \in detLogoSet$ and $brandlogo_j \in logoSet$.

3.3. Copyright feature extraction

Compared with logo and favicon, almost all copyright notices are text information. That is, there is no semantic gap for copyright recognition. And does this mean that it is easy to extract and recognize copyright notices? Actually, that is not quite right: copyright notices recognition is relatively easy, but the copyright string extraction faces a multiplicity of diverse situations. Taking © as an example, ⊙, ⊗ and ♡ are all used by phishers. Moreover, quite a number of copyright notices of phishing sites do not contain any aforementioned symbols. Figure 5 shows the screenshot of a typical phishing page targeting at Lloyds TSB. The page not only does not contain © but also does not contain "copyright" string. Even so, footer of the page still shows some copyright notice to some extent, for all the anchors linking to the legitimate Web page of Lloyds TSB.

In this paper, instead of searching "©," "&copy," "copyright," "all rights reserved" or "privacy policy," we use vision-based page segmentation algorithm to locate the block in the bottom of the page as copyright candidates [28], which is just as logo detection. The reason for this is that page footer is a block that is separated from the main body of content, where copyright notice and other statements, such as privacy policy and legal agreements, are usually presented.

After locating the footer block, we extract all the text and hyperlinks to compare with the brand strings.[§] The copyright notice feature is formulated as follows:

$$F_{copyright} = \max_i Frequency(brand_i) \quad (4)$$

where $brand_i$ is an element of brand set. *Frequency* ($brand_i$) computes the frequency of $brand_i$ strings in the footer block.

4. BRAND AUTHORIZATION FEATURES EXTRACTION

Section 3 describes in detail how to locate and represent brand features. However, the proposed features do not consider whether the Web page that embedded brand entities belongs to the corresponding brand or has an authorization to use the brand entities. In this paper, *brand authorization* refers to a one brand enterprise that acknowledges a site using its brand identity. Wang *et al.* realized the brand authorization problem in logo detection [20]. They suggested defining a new Domain Name System (DNS) record type or having brand holders embed a digital signature in their logos. The suggestion sounds interesting, but it goes beyond phishing detection itself. In this section, we propose a brand authorization features ($F_{authorization}$) to reduce the FPR of phishing sites detection.

In this paper, several features are used to judge brand authorization, which take into account domain name resolution, page redirection and incoming links information of the detected sites. One natural question arises: why these features have discriminability? To answer this question, we provide detailed illustration in the succeeding text.

Name server (NS) feature and resolution IP feature: Generally, domain resolution information explains what name servers the host uses and which IPs the server addresses. By analyzing the phishing brands in PhishTank, we find that almost all the brands have more than one domain name and that all the brands have their own domain name servers. That is, they control the domain name resolution themselves. Given these facts, we will compare the domain name servers and resolution IPs of the corresponding domain name of the suspicious URL with the brand domain names. The name servers and resolution IPs can be accessed via Domain Information Groper (DIG) services [32]. If the intersection of name servers of suspicious domain name and brand domain name is not empty, $F_{ns} = 1$; else, $F_{ns} = 0$. Instead of comparing the resolution IPs directly, we compare the prefix of IPs. The reason for comparing prefixes is that a company usually owns a range of consecutive IP addresses. We find that all domains of the popular phishing targets in Figure 3 use IPv4 addresses. So we only consider IPv4 addresses. If the intersection of the resolution IPs' prefix of suspicious domain name and brand domain name is not empty, $F_{ip} = 1$; else, $F_{ip} = 0$. In the experiments, we choose the first 24 bits of IPs. An example illustrating the effectiveness of the IP information is as follows: The "https://www.asia.hsbcprivatebank.com," called hsbcprivatebank, is a legitimate site, which is HSBC affiliated. The proposed favicon feature-based detection method will indicate that hsbcprivatebank is a suspicious phish. The host does not use the same name servers as hsbc.com.uk. That is, NS features cannot filter the URL. In this case, IP feature can help. The resolution IPs' prefix of hsbcprivatebank and hsbc has an intersection—"203.112.92."

Redirection feature: Redirection information can tell whether a Web page redirects to the brand sites. In this paper, we focus on Canonical Name (CNAME) resource record and 301 URL redirection. The former is a type of resource record in the DNS that specifies that the domain name is an alias of another [33], and the latter is the most efficient and search engine-friendly method for Web page redirection [34]. JavaScript redirection is not taken into account, since there is no reason for a legitimate brand authorization using spamdexing techniques [35]. Redirection feature can tell whether a site redirects to the brand site. A suspicious URL redirecting to brand URL means that it is highly possible that the site is a legitimate site. If so, $F_{redirection} = 1$, or else, $F_{redirection} = 0$.

Incoming link feature: The incoming links of a Web page may be of significant personal, cultural or semantic interest: they indicate who is paying attention to that page. If a detected site has incoming links from the brand sites, it indicates that the brand endorses the site. In this case, $F_{inlink} = 1$, or else, $F_{inlink} = 0$. Alexa Internet provides query interface for the in-links of a host. In the experiments, we query and extract the incoming links information via Alexa service to check whether the suspicious Web page is supported by the corresponding brand.

Some well-known phishing-targeted brands have their own autonomous system (AS), such as Paypal, Google and Yahoo. Generally speaking, AS often explicitly tells who owns an IP block. In this paper, AS is not extracted as brand authorization feature is based on the following considerations: many phishing targets do not have their own AS. For example, www.taobao.com—the most targeted phishing site—shares two public autonomous systems (AS4134 and AS4837) with many other Web sites. The two autonomous systems belong to different Internet service providers (ISPs) in China, which serves many different

[§] The brand strings include the brand name, alias, abbreviation and domain names.

Web sites, including phishing sites as long as they pay. As the main phishing targets in China, ICBC and Tencent both use the public autonomous systems belonging to different ISPs, just as Taobao does. Generally, an AS has several IP blocks, and the DNS A Record (address record) of a host name rarely changes, so choosing IP mask instead of AS can ensure a high detection precision in statistical sense. In this paper, we chose IP mask, which is based on an analysis on a variety of samples.

5. EXPERIMENTAL EVALUATION

To test the validity of the proposed anti-phishing method, we carried out experimental evaluations on a data set with many confusing cases. In the experiments, five times fivefold cross validation is run on the data set. The standard TPR, FPR, area under receiver operating characteristic curve (AUC) and F1-measure are used as the evaluation metrics.

5.1. Learning algorithm and detection features

The learning algorithm we used in the experiment is bagging, a famous meta-learning algorithm, which is widely used in Web information detection [36]. The weak classifier for bagging is C4.5. The iterations of bagging is 90 in our experiments.

Table II. Information of legitimate samples.

Source	Size	Collecting Method
DMOZ	2866	Crawling the pages of auctions, banks, payments, multiplayer games, merchant services, insurance companies and forex resources
DNS log	8330	Crawling the hosts containing brand names
Google	6375	Obtaining the URLs by querying the names of 87 most phished brands in Google, and crawling the pages

DNS, Domain Name System.

In our experiments, we extract one favicon feature, one logo feature, one copyright notice feature and four brand authorization features. Linear fusion is used for the feature fusion of proposed visual brand features and brand authorization features. Normalization of 0–1 is used to bring all values into the range $[0, 1]$ when we were training the bagging classifier.

5.2. Data set

Our Web page corpus consists of 3863 phishing cases from PhishTank and APWG, and 17 571 legitimate Web pages from three sources. The phishing samples cover 87 most phished famous brands. All the phishes are the PhishTank and APWG data from 20 February 2013 to 17 March 2013, which were crawled every day when they were alive. The phishes samples contain 15 different languages, including English, French, German, Portuguese, Chinese, Japanese, Spanish, Italian and more, which also contains 51 picture-in-picture phishes.

Similar to the previous work [3], we pay more attention to popular sites and most phishing target sites when collecting legitimate samples. Many hard cases from search engine were collected by querying the names of the most phished brands. Meanwhile, a significant amount of confusing samples were collected from DNS resolution log. Given the evaluation of the proposed method on these hard cases, the pessimistic performance statistics will be achieved, which we believe will enable the proposed method to be more convincing.

Table II tabulates the detailed data sources and methods of collecting legitimate samples. The data sources include DMOZ directory [37], DNS resolution log and Google.

5.3. Experimental results

To extract the proposed features, we collected the brands' URLs, favicons, logos and copyright notice strings. The data collection process is automated. We first obtained 158 most targeted phishing brand names from PhishTank. Then, we obtained their corresponding URLs via Google and Alexa. Next, we collected the favicons, logos and copyright notices by analyzing their home pages. This

Table III. Comparisons of phishing detection performances with different features.

Features	TP	FP	F1-measures	AUC
$F_{favicon}$	0.789	0.035	0.81	0.871
F_{logo}	0.856	0.035	0.849	0.906
$F_{copyright}$	0.834	0.034	0.838	0.895
$F_{favicon}$ and F_{logo}	0.951	0.041	0.891	0.952
$F_{favicon}$ and F_{logo} and $F_{copyright}$	0.989	0.042	0.908	0.973
$F_{favicon}$ and F_{logo} and $F_{copyright}$ and $F_{authorization}$	0.988	**0.0009**	0.992	0.993
CANTINA+	0.951	0.015	0.942	0.967
Proposed features and CANTINA+	**0.992**	0.0012	**0.993**	**0.994**

TP, true positive; FP, false positive; AUC, area under receiver operating characteristic curve.

process is easy to implement. Based on the collected data, brand features and brand authorization features are extracted for all the 21 434 samples.

On the basis of the previous work, five times five-fold cross validation is run on the data set. Table III shows the performance of phishing detection with different extracted features, where $F_{authorization} = F_{ns} \bigcup F_{ip} \bigcup F_{redirection} \bigcup F_{inlink}$. We further compare the effectiveness of the proposed features with the features used in CANTINA+ [5]. CANTINA+ is a feature-rich machine-learning framework for detecting phishing Web sites, which include 15 different features. In this table, the bold data are the best results on different evaluation metrics.

It is noticed that favicon, logo and copyright features all have good discriminability. The learnt model via brand identity features—favicon, logo and copyright features together—obtains a good performance on true positive, F1-measure and AUC measures. However, when just using brand features, the false positive is a little high, which almost kills all the legitimate brands' sites wrongly. As expected, the brand authorization features can effectively reduce the FPR. It can be observed from Table III that the brand features and brand authorization features are complementary to phishing detection.

It can also be observed that our proposed features are more effective than CANTINA+ features. That is, the data set contains many hard cases, which are undistinguishable for CANTINA+ features but recognizable for our proposed features. For example, *https://www.paypal-labs.com/devblog/* is wrongly classified as a phish by CANTINA+; nevertheless, the proposed brand authorization features can correctly recognize it as a legitimate site because it uses the same name servers as *paypal.com*.

In the experiments, we further fused the proposed features with CANTINA+ features. The model trained on all features achieves the best results on TPR, F1-measure and AUC, with a slight increase in FPR. That is, the proposed features and CANTINA+ features are complementary to phishing detection to some extent.

6. CONCLUSION AND FUTURE WORK

The core idea of the paper is to aim at the essence of phishing sites—brand spoofing, where favicon, logo and copyright notice as the most important identities of brand are widely used by phishing criminals to trick victims. In this paper, favicon, logo and copyright features are extracted first, and then redirection, incoming links and DNS resolution information-based brand authorization features are further extracted to discriminate the sites with branding rights from phishing sites. To validate the proposed phishing detection method, we constructed a big data set that contains quite a few hard cases collected from Google, DMOZ and DNS resolution log. The experimental results on the data set show the effectiveness of the proposed method. The proposed phishing detection method is a beneficial complement to the existing anti-phishing research.

The future work involves the following: (i) recognize favicons and logos via more scale-independent information such as texture features, Scale-Invariant Feature Transform (SIFT) feature and shape features, and (ii) try constructing a big enough phishing data set via crowdsourcing, and carry out more experimental evaluations on the data set.

ACKNOWLEDGEMENTS

This paper is supported by grants from the National Natural Science Foundation of China (Nos. 61005029, 61375039 and 61103138). This paper is the result of many beneficial discussions with one of our collaborators at a major Internet security firm, who wishes to remain anonymous.

REFERENCES

1. Hong J. The state of phishing attacks. *Communications of the ACM* 2012; **55**(1): 74–81.
2. Whittaker C, Ryner B, Nazif M. Large-scale automatic classification of phishing pages, In *Proceedings of 17th NDSS*, San Diego, California, USA, 2010.
3. Dhamija R, Tygar JD, Hearst M. Why phishing works, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* ACM, Montreal, Canada, 2006; 581–590.
4. Sheng S, Wardman B, Warner G, Cranor L, Hong J, Zhang C. An empirical analysis of phishing blacklists, In *Sixth Conference on Email and Anti-Spam (CEAS)*, California, USA, 2009.
5. Xiang G, Hong J, Rose CP, Cranor L. Cantina+: a feature-rich machine learning framework for detecting phishing websites. *ACM Transactions on Information and System Security (TISSEC)* 2011; **14**(2): 211–228.
6. Ma J, Saul LK, Savage S, Voelker GM. Beyond blacklists: learning to detect malicious Web sites from suspicious URLs, In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, Paris, France, 2009; 1245–1254.
7. Abu-Nimeh S, Nappa D, Wang X, Nair S. A comparison of machine learning techniques for phishing detection, In *Proceedings of the Anti-phishing Working Groups 2nd Annual eCrime Researchers Summit,* ACM, Pittsburgh, Pennsylvania, USA, 2007; 60–69.
8. Bian K, Park JM, Hsiao MS, Belanger F, Hiller J. Evaluation of online resources in assisting phishing detection, In *Ninth Annual International Symposium*

on Applications and the Internet, IEEE, Seattle, USA, 2009; 30–36.
9. PhishTank. Statistics about phishing activity and PhishTank usage, 2013. (Available from: http://www.phishtank.com/stats.php) [Accessed on April 2013].
10. APWG. The Anti-Phishing Working Group, 2012. (Available from: http://www.antiphishing.org/) [Accessed on August 2012].
11. Kumaraguru P, Rhee Y, Acquisti A, Cranor LF, Hong J, Nunge E. Protecting people from phishing: the design and evaluation of an embedded training email system, In *SIGCHI Conference on Humanfactors Incomputing Systems,* ACM, San Jose, California, USA, 2007; 905–914.
12. Downs JS, Holbrook M, Cranor LF. Behavioral response to phishing risk, In *Proceedings of 2nd Annual eCrime Researchers Summit,* ACM, Pittsburgh, Pennsylvania, USA, 2007; 37–44.
13. Fette I, Sadeh N, Tomasic A. Learning to detect phishing emails, In *Proceedings of the 16th International Conference on World Wide Web,* ACM, Banff, Alberta, Canada, 2007; 649–656.
14. Ethical Hacker. New phishing attack spread by twitter direct message, 2012. (Available from: http://www.livehacking.com/2011/07/11/new-phishing-attack-spread-by-twitter-direct-message/) [Accessed on August 2012].
15. Chloe Albanesius. Gaming apps increase spam, phishing by 50 percent, 2012. (Available from: http://www.pcmag.com/article2/0,2817,2362134,00.asp) [Accessed on October 2012].
16. Grant Gross. Phishing scam calls on VoIP phones, 2012. (Available from: http://www.pcworld.com/article/126373/phishing scam calls on voip phones.html) [Accessed on December 2012].
17. Jakobsson Markus, Myers Steven. *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. John Wiley & Sons, 2006.
18. Xiang G, Hong JI. A hybrid phish detection approach by identity discovery and keywords retrieval, In *18th International Conference on World Wide Web,* ACM, Madrid, Spain, 2009; 571–580.
19. Dhamija R, Tygar JD. The battle against phishing: dynamic security skins, In *Proceedings of the 2005 Symposium on Usable Privacy and Security,* ACM, Pittsburgh, Pennsylvania, USA, 2005; 77–88.
20. Wang G, Liu H, Becerra S, Wang K, Belongie S, Shacham H, Savage S. *Verilogo: proactive phishing detection via logo recognition*, 2010.
21. Le A, Markopoulou A, Faloutsos M. Phishdef: URL names say it all, In *INFOCOM, 2011 Proceedings IEEE,* IEEE, Orlando, Florida, USA, 2011; 191–195.
22. Maurer ME, Herzner D. Using visual website similarity for phishing detection and reporting, In *Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts CHI EA 12,* ACM, New York, NY, USA, 2012; 1625–1630.
23. Aaron G, Rasmussen R, Aaron R. Global phishing survey: trends and domain name use in 2h 2011, 2012. (Available from: http://antiphishing.org/reports/APWGGlobalPhishingSurvey2H2011.pdf) [Accessed on December 2012].
24. Geng GG, Lee XD, Wang W, Tseng SS. Favicon C a clue to phishing sites detection, In *Proceedings of the Anti-phishing Working roups 2nd Annual eCrime Researchers Summit (eCrime2013),* IEEE, San Francisco, California, USA, 2013.
25. MSDN. How to add a shortcut icon to a Web page, 2012. (Available from: http://msdn.microsoft.com/en-us/library/ms537656(VS.85).aspx) [Accessed on December 2012].
26. Dodgson NA. Quadratic interpolation for image resampling. *Transformation Image Process* 1997; **6**(9): 1322–1326.
27. Outing S, Ruel L. The best of eyetrack iii: what we saw when we looked through their eyes. Published on *Poynter Institute (not dated).* Retrieved 2004; **20**(06): 06.
28. Cai D, Yu S, Wen JR, Ma WY. VIPS: a vision based page segmentation algorithm. *Microsoft Technical Report, MSR-TR-2003-79*, 2003.
29. Afroz S, Greenstadt R. PhishZoo: an automated Web phishing detection approach based on profiling and fuzzy matching. *Technical Report DU-CS-09-03*, 2009.
30. Hu MK. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on* 1962; **8**(2): 179–187.
31. Flusser J, Suk T, Zitov B, Ebrary Inc. *Moments and Moment Invariants in Pattern Recognition*, Wiley Online Library, 2009.
32. Linux. Unix command: dig, 2012. (Available from: http://linux.about.com/library/cmd/blcmdl1_dig.htm) [Accessed on December 2012].
33. Wikipedia. Cname record - wikipedia, 2012. (Available from: http://en.wikipedia.org/wiki/CNAME_record) [Accessed on December 2012].
34. Google. *Webmaster tools help*, 2012. (Avilable from: http://support.google.com/webmasters/bin/answer.py?hl=en&answer=93633).
35. Chellapilla K, Maykov A. A taxonomy of JavaScript redirection spam, In *Proceedings of the 3rd*

International Workshop on adversarial Information Retrieval, ACM, Banff, Canada, 2007; 81–88.

36. Castillo C, Donato D, Gionis A, Murdock V, Silvestri F. Know your neighbors: Web spam detection using the Web topology, In *Proceedings of the 30th International SIGIR Conference on Research and Development in Information Retrieval,* ACM, Amsterdam, Netherlands, 2007; 423–430.
37. ODP. ODP - open directory project, 2012. (Available from: http://www.dmoz.org/) [Accessed on December 2012].

DOI: 10.1002/sec