A NOVEL METHOD TO PREVENT PHISHING BY USING OCR TECHNOLOGY

Yunjia Wang
School of Computer Science
University of St Andrews
St Andrews, United Kingdom
yw43@st-andrews.ac.uk

Ishbel Duncan
School of Computer Science
University of St Andrews
St Andrews, United Kingdom
immd@st-andrews.ac.uk

***Abstract*—Phishing is one of the most common attacks in the world, especially with the increasing usage of mobile platforms and e-commerce. Although many users are weary of phishing attacks from suspicious paths in the URL address, phishing still accounts for a large proportion of all of malicious attacks as it is easy to deploy. Most browser vendors mainly adopt two approaches against phishing; the blacklist and the heuristic-based. However, both have related limitations.**

In this paper, a novel method was proposed to protect against phishing attacks. By using image recognition (OCR) technology, phishing attacks can be distinguished from the actual website by reading the logos on the website and comparing with the site URL. An easy to implement prototype demonstrated a high accuracy of detection in the experimental trials.

***Index Terms*—Phishing, OCR, Phishing Prevention**

I. Introduction

The internet has become an indispensable element in our daily life; it provides significant resources to people whether for play, work or education. Moreover, with the increased universality of mobile devices, a magnitude of services is at our fingertips, including financial transactions. As shown by existing data from the British Bankers Association and Ernst & Young (EY) in 2017 [1], consumers are increasingly shifting to managing their assets by using mobile banking apps, and this number has continually increased over the past year, as shown in Figure 1.

Meanwhile, users still face a large challenge, which is mobile security. Both Apple and Google have different strategies to protect the security of devices and apps [2], but phishing is still hard to prevent. In a survey shown from McAfee in 2015, approximate 97% of consumers were unable to identify phishing emails correctly [3]. Moreover, with the increase in usage of QR codes, QR phishing presents a threat to this new and convenient technology [4]. Today, the majority of users are wise to phishing attacks from suspicious paths in the URL address. But different to laptop or PC, mobile systems and browsers often lack secure identification, so the user cannot easily identify the phishing URLs from the target address [2]. For phishing emails and QR phishing, the malicious URL are often made invisible, and most unwary users do not check the accessed address [5]. As a survey from Wandera shows, a new phishing site is created every 20 seconds [6]. Problematically, list-based phishing protection services are not

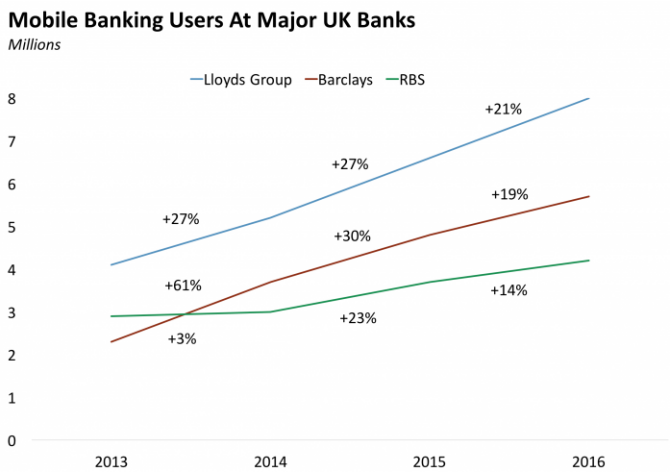

Fig. 1. Mobile Banking Users at Major UK Banks [1]

effective to isolate these suspicious sites because sites are not detected and registered in real-time [6]. According to the Kaspersky Lab anti-phishing investigation, over 246 million user attempts to access different kinds of phishing pages have been detected, with 54% being attempts to access a financial-related website versus 47% in 2016. This is the first time that financial phishing accounts for over 50% of all phishing attacks [7]. Therefore, with this increasing trend, phishing defences whether in mobile platforms or PC platforms needs to be better and faster. In this paper, we present a novel method to protect against phishing attacks and also demonstrate a high accuracy rate through several evaluation experiments.

The remainder of this paper is structured as follows: in section 2, we review the related work. In section 3 we describe the methodology used and subsequently, the experimental results are evaluated and related limitations are analyzed in section 4. Lastly, section 5 summarizes this work and includes our conclusions.

II. Related Work

We focused on three aspects in the literature reviewed: phishing, phishing prevention & detection and security strategies on mobile platforms.

In Felt and Wagner [2], the risk of phishing on mobile platforms was assessed. The concept of control transfer was introduced, which is that web sites in mobile browsers should obey the standard Same Origin Policy [1] as they are potentially untrustworthy, which indicates that web sites on different domains should be isolated from each other. However, neither Android nor Apple restrict access to mobile web sites. Thus phishing always occurs during a control transfer, such as a trusted inter-application link which may point to a malicious target instead.

In Krombholz et al. [8], the emerging technology of QR codes was shown to be a phishing attack vector as they are cheap to produce and easy to deploy. Either the attacker replaces the entire QR code or the attacker modifies a few pixels of a QR code to be used as the attack vector. These encoded malicious QR codes redirect the user to a phishing scam. As shown by existing data from Sharma [9], the first malicious case using QR codes was detected in September 2011.

The battle against phishing has been continuing and the relevant current prevention and detection strategies are presented next. Generally, most browser vendors adopt two approaches against phishing: blacklist and heuristic-based [10]. In the first method, the target URL will be checked from the phishing blacklist before accessing the URL, but this static method cannot prevent phishing completely. This is because all of the phishing URLs in the blacklist need people to report them, and there is no way to improve this blacklist dynamically, so newly created phishing websites will not exist in this pool. In the second method, phishing can be detected through the characteristics of the target, such as the content of the web site [11]. Some machine learning approaches are derived from this method.

Most machine learning methods use lexical and host-based analysis to categorize the features. For example, in some research [8] [12] [13], the textual properties in the URL are considered for lexical analysis. For host-based analysis, the server properties are investigated from WHOIS[2], such as IP address, registration information etc. However, gathering this information may become a hard problem due to possible restrictions in the future. WHOIS has continually sparked controversy as it generates too much private information. The BBC in May 2018 stated that most of this information have been wiped from WHOIS in order to comply with the EU's General Data Protection Regulation (GDPR) legislation [14].

In addition, research into security indicators of phishing on the browser is relatively mature. In the investigation by Egekman. S, Cranor. L and Hong. J [15], an active warning was more effective than a passive warning, as the user often ignores the pop-up message.

[1]Same Origin Policy: An import concept in web security. Under this policy, scripts can only access data from sites which are the same as the origin page. An origin is defined as a combination of URL scheme, host name and port number [21]. This means it ensure that the script running on a site can only access data from the origin website itself.

[2]WHOIS: Finds information on any domain name or website. https://who.is/

However, phishing prevention on mobile platforms is more complicated than expected. Besides the same problems as the desktop computer, it still faces two additional challenges. Firstly, the majority of phishing links come from phishing emails, and mobile platforms do not support secure identification [2]. The mobile user is unable to know whether the accessed URL address is safe, especially if the user lacks security awareness. For example, Google Chrome provides much better security against phishing than other web browsers [10]. It shows a warning page if the accessed URL is malicious, but the Chrome browser does not provide this service on the mobile platform, as shown in Figure 2. In addition, the user may still fall victim to a scam even if they check the URL. For example, some phishing scams exploit JavaScripts scrollTo() function to hide the original URL bar on the page, which is replaced by a fake address bar within the page [16]. This is below the real address bar to confuse the user. Although this risk has already been fixed recently, it still exists in older versions of the platform.

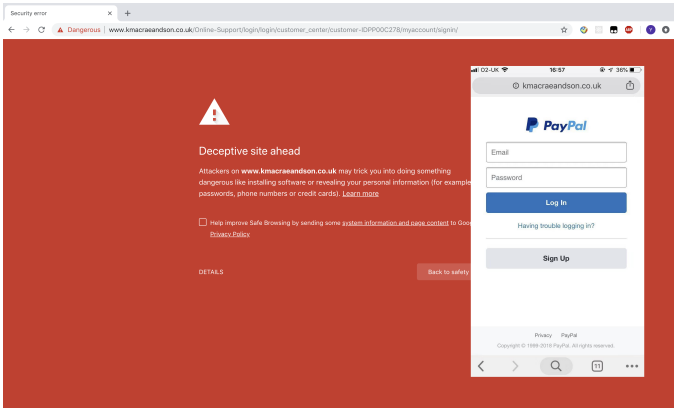

Fig. 2. Chrome access same website PayPal by using PC and mobile [10]

Secondly, there are phishing risks from QR (Quick Response) codes on mobile platforms. A QR code is encoded, so it is unreadable by humans and specific QR readers are needed to parse the information. Thus it is possible to trigger a potential vulnerability, such as buffer overflows or command injection by reading the manipulated QR code [17] [18]. Moreover, 31 QR scanner applications were compared in [19], with only two apps having a security warning feature, but they also had a higher ratio of false negative errors. Subsequently, two better open source APIs (Google safe browsing, phishtank) were recommended to improve the phishing prevention accuracy (True Positive), but the static limitations (of blacklists) has not yet been overcome.

Therefore, in this paper, we present a novel approach to defend against phishing. By using OCR[3] technology, a dynamic detection method can be deployed using the image content of the web. It not only solves the blacklist's static

[3]OCR(Optical Character Recognition): The Vision API can detect and extract text from images. https://cloud.google.com/vision/docs/ocr

limitation issue, but also avoids the privacy restriction issues from WHOIS by using related machine learning.

III. Methodology

A. Research Aim

The aim of this research is to determine the purpose of the website using the logo or the background image by applying image recognition techniques. Then, by comparing these facets to the official website, we identify potential phishing activity from the accessed URL. The implementation is divided to four steps, as shown below:

1) Extract the image
2) Describe the image content
3) Confirm the official URL
4) Verify the accessed URL

B. Preliminary Requirement

In order to verify the feasibility of the preliminary methodology, a python program was written to detect images and search related official website. As part of the implementation, some open source APIs need to be registered:

- Google Optical Character Recognition (OCR) API
- Google Search API

C. Procedures

The following steps were applied to 40 URLs taken from PhishTank and analyzed as a proof of concept.

1) Extract the image: In order to confirm the purpose of the website, the first step was to use a web crawler to extract the related images. Generally, there are two methods to link the logo or background image in most websites; HTML code and .css files. In the first approach, the logo image is usually inserted under the <img>tag in HTML code, such as Google.com or Microsoft.com. In the second approach, the logo image has been linked via .css file, under the attributes of background or background-img, such as occurs in PayPal.com. All of these conditions should be considered when we extract the background image due to the complexity and diversity of phishing.

2) Describe the image content: In this step, we use the Google Optical Character Recognition (OCR) API to detect image based content information. Because this API only works for recognizing specific text in an image, the content of redundant images, such as a symbol icon from the .css file, scenery or character images from HTML <img>tags, will be ignored. This increases the efficiency for subsequent manipulation. We also tried other OCR APIs, such as Microsoft Azure, to exam the extracted content. But the result was not as good as expected, this will be mentioned in the next section.

3) Confirm the official URL: From the description of the image content above, we can identify the expected purpose of this website. In this step, the related official URL address will be obtained by using a keyword search in the Google Search API. According to these keywords, the related official URL addresses can be confirmed, and we only kept the top three results from the returned result list in our initial experiment. Generally, the first URL may be the official URL of this website; the second may be the Wiki link about this website; and the third may be the related news about this website, or other branches of this website. For example, in one of our experiments, "PayPal" text was recognized from a phishing PayPal website, and the related official URLs were confirmed by using a Google Search from this text, as shown in Figure 3.

```
The related official url is:
{'url': 'https://www.paypal.com/us/home', 'Organization': 'PayPal, Inc.'}
{'url': 'https://itunes.apple.com/us/app/paypal-mobile-cash/id283646709?mt=8',
'Organization': 'Apple Inc.'}
{'url': 'https://www.paypal.com/login', 'Organization': 'PayPal, Inc.'}
```

Fig. 3. Search key words PayPal, and related result

4) Verify the accessed URL: As most websites have already used SSL (Secure Sockets Layer) to improve the security of the transmission due to its encrypted link, this ensures all data are private and integral under an established channel between a web server and a browser [20]. Therefore, we verify the security of the accessed URL by comparing the SSL certificate information.

First, we attempted to retrieve the SSL and associated hash thumbprint to confirm the consistency between the accessed URLs and the official URL. However, the result is not as good as we expected. The hash value may be inconsistent under different domain names or branches in the same company. For example, as shown in Figure 4, the SSL certificate hash value is different between https://www.google.com/ and https://www.google.co.uk/. So, finally, we used the organization name in SSL certificates to confirm that these websites belong to the same company, and verify the security of the accessed URL against all registered URLs.

```
{'url': 'www.google.com', 'hash': '8512c2a42dac995fa6ca65843ec38fd9'}
{'url': 'www.google.co.uk', 'hash': 'b3c781e6c93a646f70e3f26e5c831bfe'}
```

Fig. 4. SSL certificate hash comparison

IV. Evaluation

A. Analysis and Result

In order to detect the accuracy of this approach, several malicious URLs were detected from PhishTank, and all of these URLs satisfy the following conditions:

- These URLs have to be online during the detection phase.
- Their logo must contain text, otherwise an API would not return the description.
- Their logo must be in English, as the OCR API only works in English in the prototype.

In the preliminary testing, we randomly collected 40 different phishing URLs from PhishTank. We found most of the reported phishing URLs in this blacklist were about financial accounting with 72.5% (29/40) sites being financial sites such as PayPal, Alibaba, American Express, NatWest, etc., as shown in Figure 5. In this 72.5%, the highest proportion was PayPal, at 79% (23/29), as shown in Figure 6.

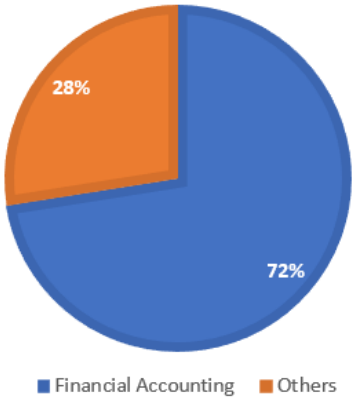

Fig. 5. 72.5% of 40 randomly chosen phishing URLs are about financial accounting

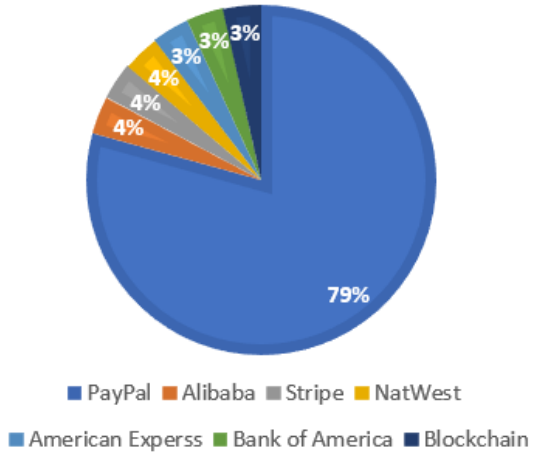

Fig. 6. PayPal was the highest proportion of 40 financial accounting sites at 79%

Under our approach, 90% (36/40) phishing URLs were successfully identified, and only 4 URLs were missed. We analyzed these failures and the main reason was due to the logo extraction phase. This means the purpose of the website in Step 2 would be incorrect if the logo image was extracted unclearly or incorrectly in Step 1. When we further analyzed the results, the specific reasons were:

- The logo was presented in text rather than by image, so it is quite hard to clearly identify this by web crawler.
- The whole page was presented in an image, so there was too much detail for the OCR API to process.

B. Limitations and Challenges

Undoubtedly, OCR technology is the crucial part in this method and the accuracy of the final result is subject to recognized details. At first, we tried to use Microsoft (Azure) OCR API in our approach. However, the testing result was not as good as expected, because it would not work under some special cases, such as a logo with a dark background colour. Also it only supported the JPEG, PNG, GIF and BMP image formats. Some websites have used SVG files as the logo image format, thus, we switched to Google OCR API in our implementation, which overcomes all the limitations of the former approach.

The recognized result is also affected by the web crawler because the recognized result is worthless if the extracted logo image is incorrect. Therefore, according to all of these issues, the limitations and challenges are concluded as:

- **The accuracy of logo image extraction**: In various versions of phishing, a logo may be inserted by a different method, it is therefore hard to locate a logo image if, for example, the whole page is an image.
- **The cost of OCR API**: There are various APIs which can be used to recognize the text from the image. However, all of them are not free. There are a free quota per day, around 1000 times, but any extra checks need to be charged for. Thus if you want to process on a larger level, the cost of OCR API should be considered.
- **The efficiency of system**: Phishing websites can be varied and complex; a phishing URL may store a lot of images, or .css files. If we have this situation, the computing process may be much longer, either for the web crawler, or the image recognition process (as all of images need to be recognized).
- **Single detectability**: The threat of phishing is varied, it can not only steal the victim's private information, but also execute a virus to infect the victims causing them to become a component of a botnet. However, using this method, it is not possible to prevent the risk from the implanted virus in a phishing site.

V. Conclusion and Future Work

Phishing attacks are always cheap to produce and easy to deploy, and most vendors have been using different approaches to prevent phishing. However, these solutions cannot keep up with the constant updating of phishing websites. In this paper, we reviewed related literature about phishing attacks and preventions, and we presented a novel approach to identify phishing websites by using an OCR technique. Unlike previous research, our approach overcomes the limitations of current methods and research solutions, not only providing a dynamic detection method, but also avoids privacy restriction issues, such as WHOIS results in machine learning. Even if the phishing server has been compromised, it can also be identified. Although this technique has a few limitations to be improved, it enables a high detection accuracy rate and the evaluation results look promising.

We aim to implement this approach in mobile platforms. On mobile platforms, more challenges would be faced, such as:

- Frequent detection may affect the limited resources on mobile platform, either in power or in network bandwidth.

- Whether it is possible to use a minimal data or no network data to detect phishing.
- How to insert this functionality in current browsers, implement all manipulations on server side to reduce the resource consumption on the mobile user side.

Therefore, we will attempt to overcome current limitations and deploy a more suitable solution to these mobile security challenges.

REFERENCES

[1] A. Aouad, Mobile banking is on the rise in the UK - Business Insider, 2017. [Online]. Available: http://uk.businessinsider.com/mobile-banking-is-on-the-rise-in-the-uk-2017-6. [Accessed: 17-Nov-2018].

[2] A. P. Felt and D. Wagner, Phishing on Mobile Devices. in Web 2.0 Security and Privacy Oakland, California, 2011.

[3] S. Cook, 50+ Phishing Statistics, Facts and Trends 2017-2018 — Comparitech, 2018. [Online]. Available: https://www.comparitech.com/blog/vpn-privacy/phishing-statistics-facts/#gref. [Accessed: 17-Nov-2018].

[4] T. Vidas, E. Owusu, S. Wang, C. Zeng, L. F. Cranor, and N. Christin, QRishing: The Susceptibility of Smartphone Users to QR Code Phishing Attacks, Springer, Berlin, Heidelberg, 2013, pp. 5269.

[5] B. Regls, Phishing News: QR Code Phishing Scheme, International Journal of Computational Intelligence and Information Security, May 2012 Vol.3,No.5, 2016.

[6] Wandera, Mobile Phishing Report 2018, 2017.[Online]. Available: http://go.wandera.com/rs/988-EGM-040/images/Phishing%20%282%29.pdf

[7] Kaspersky, Financial phishing accounts for over 50% of all phishing attacks fo..., 2018. [Online]. Available: https://www.finextra.com/pressarticle/72837/financial-phishing-accounts-for-over-50-of-all-phishing-attacks-for-the-first-time. [Accessed: 17-Nov-2018].

[8] K. Krombholz, P. Frhwirt, P. Kieseberg, I. Kapsalis, M. Huber, and E. Weippl, QR Code Security: A Survey of Attacks and Challenges for Usable Security, Springer, Cham, 2014, pp. 7990.

[9] V. Sharma, A Study of Malicious QR Codes. .

[10] N. Mazher, I. Ashraf, and A. Altaf, Which web browser work best for detecting phishing, in 2013 5th International Conference on Information and Communication Technologies, 2013, pp. 15.

[11] Y. Zhang, J. Hong, and L. Cranor, CANTINA: A Content-Based Approach to Detecting Phishing Web Sites.

[12] M. S. I. Mamun, M. A. Rathore, A. H. Lashkari, N. Stakhanova, and A. A. Ghorbani, Detecting Malicious URLs Using Lexical Analysis, Springer, Cham, 2016, pp. 467482.

[13] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, 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 - KDD 09, 2009, p. 1245.

[14] BBC, GDPR risks making it harder to catch hackers - BBC News, 2018. [Online]. Available: https://www.bbc.co.uk/news/technology-44290019. [Accessed: 17-Nov-2018].

[15] S. Egelman, L. F. Cranor, and J. Hong, Youve been warned, in Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI 08, 2008, p. 1065.

[16] Y. Niu, F. Hsu, and H. Chen, iPhish: Phishing Vulnerabilities on Consumer Electronics.

[17] K. Peng, H. Sanabria, D. Wu, and C. Zhu, Security Overview of QR Codes. - Student project in the MIT course 6.S57, '14.

[18] P. Kieseberg et al., QR Code Security. Proceedings of the 8th International Conference on Advances in Mobile Computing and Multimedia, November 08-10, 2010, Paris, France.

[19] H. Yao and D. Shin, Towards preventing QR code based attacks on android phone using security warnings, in Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security - ASIA CCS 13, 2013, p. 341.

[20] S. Chugh, Why Google is Forcing You To Have SSL Certificate on Your Websites, 2018. [Online]. Available: https://serverguy.com/security/google-forcing-ssl-certificate-websites/. [Accessed: 17-Nov-2018].

[21] Wikipedia, Same Origin Policy. W3C.