POSTER: Phishing Website Detection with a Multiphase Framework to Find Visual Similarity

Omid Asudeh
University of Texas at Arlington
Arlington, TX, USA
Omid.Asudeh@mavs.uta.edu

Matthew Wright
Rochester Institute of Technology
Rochester, NY, USA
Matthew.Wright@rit.edu

ABSTRACT

Most phishing pages try to convince users that they are legitimate sites by imitating visual signals like logos from the websites they are targeting. Visual similarity detection methods look for these imitations between the screen-shots of the suspect pages and an image database of the most targeted websites. Existing approaches, however, are either too slow for real-time use or not robust to manipulation. In this work, we design a multi-phase framework for visual similarity detection. The first phase of the framework should rule out the bulk of websites quickly, but without introducing false negatives and with resistance to attacker manipulations. Later phases can use more heavyweight operations to decide whether or not to warn the user about possible phishing. In this abstract, we focus on the first phase. In experiments, our proposed method rules out more than half of the test cases with zero false negatives with less than 5 ms of processing time per page.

Keywords

Web security, phishing, visual similarity

1. INTRODUCTION

In the US alone, hundreds of millions of dollars are lost to phishing attacks each year [4]. Thus, despite substantial research into anti-phishing mechanisms, attackers continue to find the attacks to be successful and profitable.

While a variety of information could be used to detect phishing sites, such as DNS Whois records and IP address of the server, much of it can be manipulated by an intelligent adversary. One aspect that is harder to manipulate, however, is the visual branding that helps the attacker convince the user that the site is the legitimate target site, such as a particular online bank. Without these visual cues, or with heavily modified cues, the user may find the site to be suspicious, and the attacker's success rate will be lower.

Thus, a well-studied approach to combating phishing attacks is to look for *visual similarity* between an unknown website and popular phishing targets, like online banks and e-commerce sites. Unfortunately, prior work [6] in using visual similarity is either too slow for real-time use or not robust to attackers manipulating parts of the page.

1.1 Contributions

In our work, we first note that branding, particularly in the form of logos, is key to users' visual recognition of a site. A site can change its overall design and features, but the logo provides a stable visual reference for customers to associate with the company. Thus, while an attacker may manipulate parts of a page, the logo should be shown and not greatly changed from the original. We focus our attention, then, to finding visual similarity[1] between logo elements on the new page and the original page. Also, we seek to ensure that our methods are robust to manipulations that do not cause the logo to change dramatically from the original, such as background color shifts or small rotations of the image.

Our primary contribution is to propose a multi-step framework for finding and testing the visual similarity of logos that works in real time. The framework is made up of a pipeline of classifiers. The goal of the first classifier is to quickly and confidently rule out many of pages that are not phishing attacks so that they can leave the pipeline and not be subject to further processing. A key feature of this stage is to minimize the false negative rate, so that phishing pages will not be shown to the user without further checks. Suspect pages will be passed to further classifiers, which should be more accurate though possibly slower. Filtering out many pages in early phases will keep most pages from having to pass through slow classifiers and keep the average processing time low.

The key insight of this multi-phase framework is that a typical website does not have any part of the main screen that looks like, for example, the Paypal logo. Such a page can be quickly determined to not be imitating the Paypal page. Then, if turns out that the page has a logo that looks like the Ebay logo, more resources can be used to determine accurately if the two logos are close enough to confuse users.

2. SHALLOW DETECTION

In this section, we describe the details of the first phase of the system, the *Shallow Detection* phase. This phase is meant to eliminate a large fraction of benign pages from further inspection and thereby reduce processing time. In particular, the objectives of this phase are: 1. It should

[1]Note that an exact copy of the logo may be easily detected.

CCS'16 October 24-28, 2016, Vienna, Austria

ACM ISBN 978-1-4503-4139-4/16/10.

DOI: http://dx.doi.org/10.1145/2976749.2989050

Figure 1: Gradient images in X and Y directions

be efficient to process any page without noticeable delay for the user; 2. It should have almost no false negatives; 3. It should remove a substantial fraction of false positives; and 4. It should be robust to attacker manipulation of the images.

2.1 Image Similarity

Our approach searches for a match for any logos on the page in a database of legitimate page logos. This scenario will push us to the area of content-based image retrieval. There are several algorithms for accurate image description and detection such as SIFT [5], but their computational overhead is high, making them inappropriate for our purpose. A classic method for content-based image retrieval is to represent the image with its histograms of color or intensity, which makes for fast comparisons. Unfortunately, they can be fooled by simply changing the background color of the logo.

Histograms of Gradients. Edges convey information about the structure of the objects in the image, which is missing from color histograms. Dalal et.al. [2] introduced *histograms of gradients (HOG)* as an image representation method in a human detection context. The gradient shows the amount of change in the image, which is typically highest along the edges. Figure 1 illustrates the gradient images in both X and Y directions. Using a two-dimensional gradient vector for each pixel of the image, we get nine bins for direction and a magnitude for each bin. A histogram of this data is used to match with images in the database.

Histogram Similarity. Once we have histograms that we can use to match images, we must have a metric to compare the histograms with. We chose *histogram intersection* as our histogram similarity measure, which is used in content-based image retrieval [3]. Using these similarity scores, we can set a simple threshold τ, $0 \leq \tau \leq 1$ and state that any image with a similarity score greater than τ is a match.

3. EVALUATION

3.1 Data sets

We use two major data sets in our experiments. The first one is the *valid phishing logo data set*, which includes 132 logos extracted from valid phishing pages in PhishTank. More than 1000 screenshots of the phishing pages, collect between Jan. 1 and Feb. 5, 2016, were examined by humans and 132 logos were extracted. The logos are not unique; for instance there are eight different phishing logos targeting Facebook in the data set. The second data set is *the legitimate logo data set*, which includes 168 legitimate logos of 65 mostly targeted brands. We checked the monthly status of PhishTank for all 12 months of 2015 and selected the top 20 targeted companies. We then took the union of this set with list of the most targeted brands published by APWG in 2014. We also added the most targeted companies among the 1000

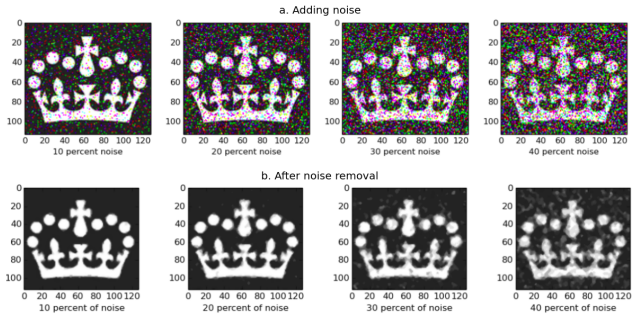

Figure 2: a. Adding noise, b. noise filtering

valid phish screenshots found in the phishing dataset. Finally, we manually extracted the legitimate logos of these brands from their official websites. In total, 168 logos are in this dataset (some of the brands have more than one logo).

3.2 Minimizing False Negatives

In our first experiment, we study parameters of our approach that would ensure low false negative rates with acceptable false positives. For each similarity threshold (ranging from 0.1 to 0.95, with a step of 0.02) and for each different segmentation (1, 4, 9, 16, and 25), we compared the whole valid-phish-logo data set with the whole legitimate-logo data set. For each setting, we calculated the false negative and false positive rates. The best setting of our system, with a similarity threshold of 0.83, had a false negative rate of 0 and a false positive rate of 47%.

The time for each malicious logo to be queried in the legitimate database in this setting (one segment, similarity threshold of 0.83) was 4.1 ms median and 4.5 ms average. These results shows that our comparison method, in the best setting, fits appropriately with the objectives of the Shallow Detection phase. Querying the legitimate data set is fast enough for real-time use, and the false negative rate is zero.

3.3 Noise, scaling, and rotation

We selected 32 legitimate logos and, in separate experiments, added to each of them several degrees of noise, scaling and rotation. We discuss each of these in turn.

Effect of noise. For this experiment we added *salt and pepper noise* from 10 up to 80 percent, with a step of 10 percent (see Figure 2.a) to 32 images and then compared them with the original images. When the HOG similarity was less than 0.83, the number of false negatives was incremented by one. Adding 10 percent noise caused more than 20 percent false negatives. To mitigate this, we applied *median filtering* [1], in which each pixel will be replaced by the median intensity value of its neighbor pixels in a 3x3-pixel block. Figure 2.b illustrates the results of median filtering. The results both using the noise filter and not including the filter are shown in Figure 3.a shows the improvement of false negative rate after the noise removal. The noise removal process had an average time overhead of two milliseconds.

Effect of scaling. In a similar experiment, we scaled 32 logos down from their original size by using a *scaling factor* by which we multiply both the x and y dimensions of the image. We used scaling factors ranging from 0.9 down to 0.1. The scaled logos are compared with the original logos

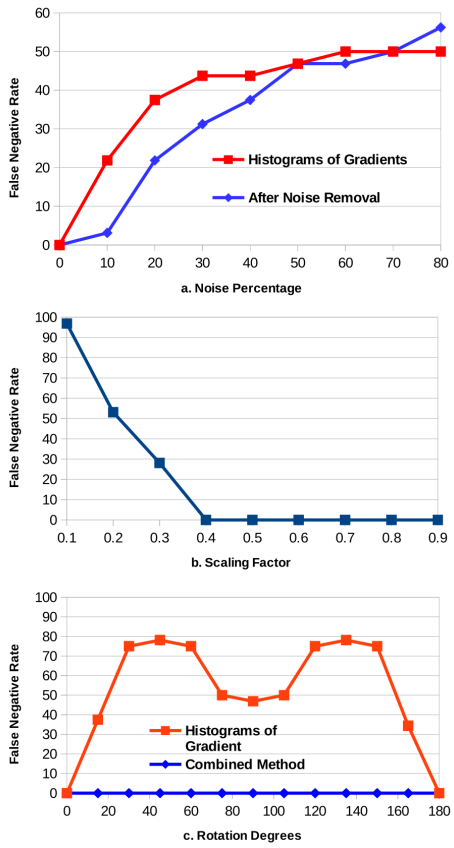

Figure 3: Effect of noise, scaling, and rotation on false negative rate

Figure 4: Scaling the logos, scaling factor 0.9 to 0.1

and if the HOG similarity was less than 0.83, the number of false negatives was incremented by one. Figure 3.b shows the results of this experiment, where the false negatives are zero for scaling factors of 0.4 and higher. As illustrated in Figure 4, scaling below 0.4 creates a qualitatively different image.

Effect of rotation. Finally, we rotated 32 logos from 15 to 180 degrees, with a step of 15 degrees and compared them with the original logos. If the HOG similarity was less than 0.83, the number of false negatives was incremented by one. We found that the similarity measure based on the logos' HOGs is sensitive to rotation. A 15 degrees rotation is enough to generate about a 40% false negative rate, and that is because rotation can dramatically change the gradients on both the x- and y-axes. To mitigate this issue, we added the intensity histograms to the similarity measure. In this case, two logos will be considered as not similar if the HOG similarity is less than 0.83 *and* the similarity of their intensity histograms is less than 0.9. We have repeated the previous experiments with the new similarity measure. This can filter the effect of the rotation completely. The results both with and without the filter are shown in Figure 3.c.

4. DISCUSSION

One of the most important assumptions in our work was that the phishing web pages visually imitate the target web pages. During the analysis of the phishing pages, we noticed a few phishing sites that asked users to provide their login credentials for a legitimate site without having any visual signal or similarity to the brand they were attacking. The method we proposed for the Shallow Detector cannot find phishing pages that have no visual imitations of the legitimate sites. Fortunately, many users can avoid being the victim of such traps, since the users have no visual basis to believe that this is the correct site. Seen another way, if our framework was successful, it would force all attackers to adopt these plain, unbranded pages and thereby reduce the effectiveness of the attack.

Future Work. Up to now our focus was on the Shallow Detector because of its importance in the overall performance of the framework. One aspect of this stage that we did not discuss is that we must be able to quickly find possible logos in the page, since the attacker may make the logo harder to find by embedding it in a larger image. We have developed a fast method to find logos based on the rendered page and will evaluate it for resilience to potential attacks. Also, since we base our design on the assumption that users are more easily fooled when the logo is present and the design of the page overall matters less, we should examine these assumptions with user studies. Next, we will explore designs for a second phase detector that works faster than existing tools and can further eliminate false positives. Finally, we will build a browser plugin that incorporates all of the phases and study it in real-world scenarios.

Acknowledgements.

This work was sponsored in part by a grant from the UT Arlington IRP program.

5. REFERENCES

[1] R. H. Chan, C.-W. Ho, and M. Nikolova. Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. *IEEE Transactions on Image Processing*, 14(10):1479–1485, 2005.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, pages 886–893, 2005.

[3] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern recognition*, 29(8):1233–1244, 1996.

[4] T. Kitten. FBI Alert: Business Email Scam Losses Exceed $1.2 Billion. http://www.bankinfosecurity.com/fbi-a-8506/op-1, 2015. Accessed: 2016-03-12.

[5] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE ICCV*, pages 1150–1157, 1999.

[6] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng. Detection of phishing webpages based on visual similarity. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 1060–1061. ACM, 2005.