PHAD- A Phishing Avoidance and Detection Tool Using Invisible Digital Watermarking

Sonali Batra
Computer Science and Engineering (CSE)
University at Buffalo (SUNY)
Buffalo, NY, USA
sonaliba@buffalo.edu

***Abstract*—This paper presents PHAD - a phishing avoidance and detection tool that uses robust invisible watermarking to watermark the logo image of a website with its domain name. The assumption behind this is that phishers copy the content of the legitimate website including the logo image. However the domain name of the attacker will be different from that of the legitimate site. On the client side, PHAD extracts the hidden watermark in the logo and compares it to the domain name. If they match then the website is deemed legitimate else a warning message appears in the browser. PHAD is intended to be a first defense and not a complete solution by itself because any watermark can be removed by manually observing the client detection process and changing the pixel values. The aim of PHAD is to significantly increase the effort required by phishers to generate an authentic looking phished website. The watermark is thus robust which is difficult to remove manually or automatically as opposed to fragile or semi-fragile watermarks which fail to be detected after benign and malignant transformations respectively.**

Keywords — phishing; robust digital watermarking; steganography; firefox extension; outguess; web browser

I. Introduction

Phishing is a form of social engineering in which black hats or hackers trick users into thinking that a fake webpage is real and into revealing their personal information in a web form like bank account details, social security number and email addresses. The black hats can then either sell this information for a lot of money in the black market or use it themselves to extract funds illegally from a victim's bank account. A large number of people are becoming victims to phishing attacks every day. According to the APWG Phishing Activity Trends Report 2013 [13], the number of unique phishing sites in September 2013 alone were 45,115. Phishers typically send a personalized email to users in which they use some form of social engineering into tricking them to visit their fake webpage. For example, they might pretend to be from a person's bank account and may urge the user to login to her account and confirm her banking activity urgently. The number of unique phishing email reports received by APWG from customers in September 2013 was 56,767 [13]. When the user logs into the fake page through her legitimate username and password, these are sent to the phisher who can then use them for malicious purposes. Phishing attacks succeed typically because *a)* some laymen are unaware that phishing attacks exist in general and *b)* Even those who are aware of them do not know how to tell a legitimate website from a fake one. Some users have an idea that they should look at the url in order to confirm that the website is real. In order to trick this class of users, phishers typically perform some play on the original domain name to use in their fake url. For example, it is not possible for an attacker to obtain bankofamerica.com domain name since it is already taken. Thus they can purchase the domain attacker.com and name the phishing page bankofamerica.banking.america.attacker.com. When a layman sees the term 'bankofamerica' in the url, she assumes that the website is legitimate, even though the actual domain name is attacker.com not bankofamerica.com. According to APWG, 56.22% of the unique phishing sites discovered in Sept 2013 contain some form of the target name in their url [13].

This paper presents PHAD - a phishing avoidance and detection tool that uses robust invisible digital watermarking to detect fake webpages. PHAD is a downloadable extension to the Firefox web browser. It is based on the fact that the domain name of a website uniquely identifies the website. In this scheme, the logo image of a web site is invisibly watermarked with the domain name of the site. When the browser visits that website, PHAD extracts the hidden message in the logo image and compares it to the actual domain name of the site. If a phisher has copied the image, the extracted message and the actual domain name shall not match as the domain name of the phished website shall be different from that of the legitimate website. Thus PHAD shall display a warning message in that case.

PHAD must be implemented using a form of robust watermarking since it must not be possible for the phisher to remove the watermark by automatically or manually distorting the image. However, if the attacker has access to a watermark detector device, she can manually remove the watermark by "experimentally deducing the behavior of the detector and exploiting this knowledge to ensure that a particular image does not trigger the detector"[14]. In doing so, she can remove the watermark by changing the values of a few pixels. However, this method is tedious and to our knowledge no known software exists to automate it at the time of writing. It can of course be automated, however that significantly increases the effort required by the attacker to generate an authentic looking webpage.

The implementation of a prototype of PHAD was successful in identifying legitimate and phished web sites based on their domain names. PHAD is implemented using outguess - a universal steganographic tool. A well known public key is used to watermark the image on the server side and detect the watermark on the client side. This avoids the problem of key exchange that would have arisen had a secret key been used. Another favorable feature of PHAD is that the logo image only needs to be watermarked once for all users versus watermarking the image separately for each user.

Topkara *et al*. [4] state that in phishing detection, "A good defense mechanism must require an integrity check mechanism that "travels with the content" when it is used or misused". Digital watermarking is one way in which this can be achieved. Through digital watermarking we can irreversibly embed some unique characteristic of a web site into resources of the site that are commonly copied by phishers such as "images, style sheets and script files", as stated by Hemanth *et al.*[10].

This paper is organized as follows. In section 2, previous work related to this research is presented. Section 3 discusses the working of PHAD. The implementation is explained in Section 4. In section 5, the threats faced by this tool or how phishers can succeed in tricking users in spite of this tool are elaborated upon. In Section 6, some questions that readers might have about PHAD are addressed. Finally in section 7, the future direction of this work is presented.

II. Related work

Huajun *et al.*[11] propose a scheme which is similar to ours. It watermarks a hash of a concatenation of several parameters including the domain name, into the source code of the website using equal tag method. The difference between this and our method is that the equal tag watermark can easily be removed if the phisher is aware of the scheme. Huajun *et al* base the security of their scheme on the assumption that the phisher is unaware of their scheme. However, in the real world algorithms are often published. Also, it is easy for the attacker to detect what the scheme is by performing experiments with a watermark detector [14]. Also, Huajun *et al* use a semi-fragile watermark which easily fails detection if malignant transformations are made to it. In contrast, we do not make any such assumption that the attacker is unaware of our scheme. Also, we use a robust watermark which is difficult to remove even by malignant transformations. Lastly, Huajun *et al.* do not address questions like what if a website has multiple domain names. We in contrast have done thorough research on how and why PHAD can fail and publish the findings in this paper.

Topkara *et al.* [4] propose an approach-ViWiD, a "visible watermarking based defense" against phishing . In this scheme, there is a "shared secret between the user and the company" [4] - a "mnemonic" [4]. The web site logo is embedded with that mnemonic and sent to the user's web browser when she visits that particular web site. If the logo does not contain the mnemonic, this implies that the site is a phished site as phished sites do not know what the shared mnemonic is. This scheme has a major drawback which is that the establishment of the shared secret between the user and the company is in itself a point of attack. It is assumed in the paper that there is "a secure connection" [4] at the time the user chooses the mnemonic in order to prevent the disclosure of the mnemonic to eavesdroppers [4]. However, this "secure connection"[4]can really not be guaranteed. A possible alternative can be to establish this shared secret out of band, however this is extremely inconvenient. And what happens if the user forgets his mnemonic or requires it to be changed? In PHAD, we use a well known public key for watermarking and detecting the watermarks in the logo images. Thus, we do not need to establish a secret key. All websites' logo images are encrypted and decrypted by the same public key. Thus the public key does not provide any security feature in PHAD.

Steel *et al.* [5], propose AIIIS-the "Automated Impersonator Image Identification System". In this approach when a phisher copies an image from a website, her server name, IP address and the date and time of the image request are digitally embedded in the image before it is served to her. Thus we can later recover this information from the image on the fake website thereby identifying the phisher. This approach is very good when it comes to identifying the phishers, however it requires prior knowledge that the site is phished. This prior knowledge comes from other anti-phishing tools or user vigilance. Thus this is not a phishing detection tool. Rather it is a phisher identification tool. Also, there may be some performance degradations when this scheme is applied to the real world scenario since it requires a unique watermark to be inserted for *each* download. PHAD is a better approach since it only requires the logo image to be watermarked once. It does not require the logo image to be watermarked separately for each user.

Other non watermarking based anti-phishing solutions are as follows. Wenyin *et al.* [7], present an approach by which a legitimate website owner can search the Web for websites mimicking his site. A website is reported as a phishing suspect if it is "visually similar" to the legitimate website [7]. This approach is advantageous for website owners however it does not directly alert the *user* whether a particular website that he is visiting is phished or not. Ronda *et a*l. [8] present iTrustPage-"a user assisted anti-phishing tool" [8]. This asks the user to describe the web form they are thinking of filling out in a few words. It then feeds these words into Google. It then compares the domain name of the suspicious website to the search results returned by Google. If the suspicious website's domain name matches any of the top 10 search results, the site is considered legitimate. The logic behind this is that Google shall presumably pull up the legitimate website and that should be within the top 10 results obtained. Even though this approach is effective, its limitations are obvious. It requires the *users'* assistance thereby inconveniencing them by utilizing their time and effort. Also it is relying on the Google search engine which can be tricked into pulling out the phishers webpage. Zhang *et al.* [3], discuss CANTINA - “a content based approach to detecting phishing web sites” [3]. First the authors explain TF-IDF, a well known information retrieval algorithm. They state that a term in a given document has a high TF-IDF weight if it is common in that document but at the same time relatively uncommon in other documents of the collection. CANTINA works as follows. It calculates the TF-IDF of each word of a website. Then it feeds the five words with the highest TF-IDF to Google. If the domain name of the suspicious web site matches any of the domain names of the top 10 results obtained, the website is considered legitimate. The logic here is that the five terms having the highest TF-IDF on a website shall pull up the legitimate website in Google since these five words are common to the legitimate website and uncommon to the entire collection of websites on the Web. Also, the phished website shall not appear in search results since it is hardly ever referred to. The limitation here is that again, the authors are relying on Google which can be tricked. Garera *et al.* [9] describe a “framework for the detection and measurement of phishing attacks”. In this work, a “logistic regression filter” is developed based on several criteria that distinguish between a legitimate and a fake URL. Chandrasekaran *et al.* [6], describe an approach in which “fake responses” are given to the website instead of real responses by legitimate users. The behavior of the website to the fake responses is recorded and fed to a decision engine which determines if the website is legitimate or phished[6].

III. How phad works

Phishers want the look and feel of their phished website to be as similar as possible to that of the original website**.** In order to achieve this, they copy the content of the legitimate website and put it in their fake website, including the *images*. PHAD uses this fact to its advantage. The logo image of the legitimate website is digitally watermarked with the domain name of the site. On the client side, PHAD compares the extracted message in the logo image to the actual domain name of the website in question. If they do not match, this implies that the website is phished. This is since the only possible way in which this could have happened is if the phisher had copied the logo image from the legitimate site. Of course, a phisher could use software in order to generate from scratch an image that looks like the logo. Thus it is suggested to make the logo sufficiently complicated and sufficiently noisy so that this process becomes tough for the attacker. Also we restate that PHAD is intended to serve only as a first defense and not a complete filter, to account for exceptionally artistic hackers who have all the time in the world.

IV. Implementation

PHAD is implemented as a downloadable extension to Firefox. It is a .xpi file (source code at acsu.buffalo.edu:/~sonaliba). PHAD works as follows. When a user visits a website, the logo image is downloaded to the hard disk. Then that image is submitted as an argument to outguess - a universal steganographic tool. Outguess detects the watermark in the image and writes it to a file on the hard disk. It uses a well known public key in order to detect the watermark. Then the file is read and its contents are stored in a variable. Next, the domain name of the website is retrieved and stored in a variable. Finally the two variables are compared. If they are equal, this implies that the website is legitimate otherwise it implies that the website is phished.

V. Threat Model

PHAD can be defeated in the following scenarios-

1) If the phisher manages to remove the watermark from the logo image she has copied from the legitimate website. Then she can re-watermark the image with the domain name of the phished website. However, it is very difficult to remove robust watermarks without severely distorting the image. One way in which this can be done is to “experimentally deduce the behavior of the watermark detector and exploit this knowledge to ensure that a particular image does not trigger the detector”[14]. In doing so, she can remove the watermark by changing the values of a few pixels. However, this is tedious and time consuming and to our knowledge no known software exists to automate this process at the time of writing.

2) The phisher can manually create a similar looking logo using some software like Paintbrush. However this will significantly increase the effort required by the phisher since this cannot be automated.

3) The phisher can observe the client detection software and take similar steps in order to remove the watermark since she now knows what pixels the watermark consists of [14]. This is a real threat since this can be automated as well. Also, the phisher need not invert the watermark exactly. She can simply remove it so that it is not visible by the naked eye even after magnifying. However, no known software like this exists at the time of writing that automates this process.

The readers are requested to keep in mind that PHAD only serves as a first defense against phishers and does not consist of a comprehensive phishing detection software. It can be used to filter out one layer of phishing websites.

4) What if a phisher takes a screenshot or photograph of the image? Since the message is digitally watermarked into the original image, it shall persist across screenshots. It shall also persist across photographs if the quality of the picture is good. If not, then its is easy for the user to detect that it is not the original image.

VI. Other Questions

In this section we have attempted to answer additional questions that the reader might have about PHAD. eg. What if a company has two or more domain names? Also what if CNN wants to run a story on Facebook and has Facebook's logo embedded in its page? Will the extension give a warning message for the CNN site too? The answers are below -

1) *What if a company refuses to watermark its logo image?* - This is a real limitation since in order for PHAD to

correctly detect phishing websites, all website owners must agree to comply with this scheme. It is however not impossible to achieve this in practice.

2) *What if a company has multiple domain names?* - For example, google.in for India and google.us for the United States. The solution would be to have all the domain names watermarked into the logo and the client will check to see if one of them match the actual domain name.

3) *What if a company has multiple logos?* For example, Twitter. In this case all the domain names of the company would be watermarked into both the logos.

4) *What if a website wants to embed a logo of another company in its page?*- For example, If CNN wants to write a story on Facebook and embed the Facebook logo in its page? Will our extension show a warning message for CNN.com which is a legitimate site? A solution to this is that there should be only one watermarked logo in a page and it should clearly be marked. A much better solution is that multiple watermarked images be allowed on a page and the company having the highest ratio of images be compared to the domain name of the site. For example, If the CNN web page has 6 images watermarked by CNN, 2 by Facebook, 1 by Yahoo and 5 unwatermarked images, the ratio of images watermarked by CNN in the page is the highest. Thus we assume that the site is pretending to be or is CNN. Thus the domain name of the site is compared to CNN.com and it would determine that the website is legitimate.

5) *What if a website has two or more watermarked images?*- If a website were permitted to have two or more watermarked images and the watermarks of all of them were compared to domain name to see if any match, an attack would be possible. This is that if the attacker put her own logo watermarked with her own domain name, and also Facebook's logo with its original watermarking. In this case PHAD would fail because it would still declare the attacker's website to be legitimate since it has a logo that is watermarked with the correct domain name. Thus instead of comparing all images' watermarks, the watermark having the highest ratio to all should be compared like in the above case.

6) *How is this better than using https?* - We claim that our method is better than using https in the following ways-

a) Users are not aware that the url they are visiting should be preceded by a https rather than http. Of course, there are browser add ons that alert the user of the fact. However, most users are not aware of the add on itself. The users might be more aware of PHAD if it were to be advertised properly.

b) HTTPS requires a central authority for certificate handling. This is not present in PHAD. Thus there is no management bottleneck. There is also no single point of failure in PHAD like in HTTPS.

c) HTTPS can be used as an added security measure along with PHAD. Like we mentioned earlier, PHAD is only intended to be a first defense against phishing and not the complete defense.

7) *What is the attacker removes and re-watermarks the image?* - It is potentially possible for the attacker to remove the watermark by observing the steps taken by the client detection software and replaying those [14]. We recommend the websites to have a noisy logo as opposed to a less noisy one so that this process becomes more difficult to the attacker. This shall increase the effort required for her to create an authentic looking phishing site.

8*) What is the purpose of the public key?*-- The public key is well known and does not provide any security feature in PHAD. It is used in this context because Outguess uses a public key in order to watermark the image and to detect it on the client side. If another watermarking algorithm that does not use a public key is used for implementation, the public key can be omitted from this paper.

VII. Future Work

Future Work in this area is to extend this scheme so that PHAD may be used independently and not just as a first defense system. This would involve finding a scheme by which it would not be possible for an attacker in possession of a detector to remove the watermark.

VIII. Conclusion

In this paper we have presented PHAD- a phishing avoidance and detection tool that invisibly watermarks the logo image of the website by the domain name of the website. If a phisher has copied the image, the extracted message from the logo image shall not match the domain name of the website. Thus we shall be able to find out that the site is phished. PHAD uses robust watermarking since it must not be possible for a phisher to remove the watermark in the logo image by applying a few simple transformations to it. PHAD uses a well known public key for watermarking and detection of the watermark, thus avoiding shared secret establishment problems. Also, in this technique the logo image only needs to be watermarked once for all users versus being watermarked differently for every user. The implementation of a prototype of PHAD has been successful in identifying legitimate and phished web sites. We have also argued how PHAD is a more effective phishing avoidance and detection tool than some of the existing tools in the 'Related Work' section. We would also like to state once again that PHAD is intended to be used only as a first defense against phishing attacks since every watermarking algorithm can potentially be observed by observing the client when it is detecting the watermark and replaying the same steps [14]. We also recommend that the websites use a noisy image so that this process becomes harder for the attacker. The aim of PHAD is to significantly increase the effort used by the attacker in order to create an authentic looking phishing website.

Acknowledgment

I am extremely grateful to Professor Nick Feamster without whose inspiration this work would not have started. I would also like to thank my colleagues Brandon Ross and Vikas Dhiman for serving as a sounding board against my ideas.

References

[1] Rachna Dhamija, J.D.Tygar, and Marti Hearst. Why Phishing Works. *In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*,pages 581-590, New York,NY, USA, 2006, ACM. ISBN 1-59593-372-7 http://escholarship.org/uc/item/9dd9v9vd

[2] Anti Phishing Working Group,Phishing Activity Trends Report-Second Half 2008. http://www.antiphishing.org/reports/apwg_report_H2_2008.pdf

[3] Yue Zhang, Jason I. Hong, and Lorrie F. Craynor. Cantina: a content-based approach to detecting phishing web sites. *In WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 639-648, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7.doi:

[4] Mercan Topkara, Ashish Kamra, Mikhail J. Atallah, and Christina Nita-Rotaru. ViWiD: Visible Watermarking based Defense against Phishing. *In Digital Watermarking. 4th International Workshop, IWDW 2005. Proceedings (Lecture Notes in Computer Science Vol. 3710)*, p 470-83, 2005. url: http://homes.cerias.purdue.edu/~crisn/papers/iwdw_2005.pdf

[5] Chad M.S. Steel, Chang-Tien Lu. Impersonator identification through dynamic fingerprinting. In Science Direct:Digital Investigation,Volume 5, Issues 1-2, September 2008, Pages 60-70. url:http://www.chadsteel.com/Publications/phishing.pdf

[6] Madhusudhanan Chandrasekaran, Ramkumar Chinchani, and Shambhu Upadhayaya. Phoney: Mimicking user response to detect phishing attacks. *In WOWMOM '06: Proceedings of the 2006 International Symposium on World of Wireless, Mobile and Multimedia Networks,* pages 668-672, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2593-8. doi: http://dx.doi.org/10.1109/WOWMOM.2006.87.

[7] Liu Wenyin, Guanglin Huang, Liu Xiaoyue, Zhang Min, and Xiaotie Deng. Detection of phishing webpages based on visual similarity. *In WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web,*pages 1060-1061, New York, NY, USA, 2005. ACM. ISBN 1-59593-051-5.doi: http://dl.acm.org/citation.cfm?id=1062868&dl=ACM&coll=DL&CFID=299199812&CFTOKEN=57391165

[8] Troy Ronda, Stefan Saroiu, and Alec Wolman. Itrustpage: a user assisted anti-phishing tool. *In Eurosys '08: Proceedings of the 3rd ACM SIGOPS/EuroSys European Conference on Computer Systems 2008,* pages 261-272, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-013-5. doi: http://doi.acm.org/10.1145/1352592.1352620

[9] Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin.A framework for detection and measurement of phishing attacks. *In WORM '07: Proceedings of the 2007 ACM workshop on Recurring malcode,* pages 1-8, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-886-2. doi: http://doi.acm.org/10.1145/1314389.1314391.

[10] Hemanth Pai, and Vinaya Shenoy. Phishing Content Analysis.CS6262 Project Report. Spring 2009. Georgia Institute of Technology. Professor: Nick Feamster. Unpublished.

[11] Huajun Huang, Yaojun wang, Lili Xie and Liqing Jiang.An Active Anti-phishing Solution Based on Semi-fragile Watermark. http://docsdrive.com/pdfs/ansinet/itj/2013/198-203.pdf

[12] Nobukatsu Takai and Yuto Mifune . Digital Watermarking by a holographic technique.http://www.opticsinfobase.org/ao/abstract.cfm?uri=ao-41-5-865

[13] Anti Phishing Working Group,Phishing Activity Trends Report-ThirdQuarter 2013 http://docs.apwg.org/reports/apwg_trends_report_q3_2013.pdf

[14] Ingemar J. Cox , Jean-Paul M. G. Linnartz :Some General Methods of Tampering with Watermarks. http://www.sps.ele.tue.nl/members/J.P.Linnartz/papers/articles/98_j2_some_methods.pdf

[15] Wikipedia- Robust Digital Watermarking http://en.wikipedia.org/wiki/Digital_watermarking#Robustness