ViWiD : Visible Watermarking based Defense against Phishing

Mercan Topkara Ashish Kamra Mikhail J. Atallah Cristina Nita-Rotaru

Center for Education and Research in Information Assurance
Computer Sciences Department
Purdue University
West Lafayette, Indiana , 47907
{mkarahan,akamra,mja,crisn}@cs.purdue.edu

Abstract. In this paper, we present a watermarking based approach, and its implementation, for mitigating phishing attacks - a form of web based identity theft. ViWiD is an integrity check mechanism based on visible watermarking of logo images. ViWiD performs all of the computation on the company's web server and it does not require installation of any tool or storage of any data, such as keys or history logs, on the user's machine. The watermark message is designed to be unique for every user and carries a shared secret between the company and the user in order to thwart the "one size fits all" attacks. The main challenge in visible watermarking of logo images is to maintain the aesthetics of the watermarked logo to avoid damage to its marketing purpose yet be able to insert a robust and readable watermark into it. Logo images have large uniform areas and very few objects in them, which is a challenge for robust visible watermarking. We tested our scheme with two different visible watermarking techniques on various randomly selected logo images.

1 Introduction

Our society has increasingly become a digital society where many critical applications and services are provided on-line. Examples of such applications are financial services, retail services, on-line news channels and digital libraries. This paradigm shift has had a beneficial effect on business and education by providing faster and easier access to services and information. Unfortunately, it has also exposed these services to malicious attacks that are more difficult to detect and defend against. One of the major security concerns in cyberspace, having impact on individuals as well as businesses and organizations, is identity theft. According to a recent Congressional Statement of the FBI Deputy Assistant Director [1], on-line identity theft represents a significant percentage of the total number of crimes committed in cyberspace.

Phishing is a form of on-line identity theft in which attackers send fraudulent e-mails and use fake Web sites that spoof a legitimate business in order to lure unsuspecting customers into sharing personal and financial data such as social security numbers, bank account numbers, passwords, etc. The incidence of phishing attacks has increased significantly over the last couple of years. By the end of December 2004, Symantec Brightmail AntiSpam antifraud filters were blocking an average of 33 million phishing attempts per week, up from an average of 9 million per week in July 2004 [2]. Acknowledging that phishing is a significant threat to e-commerce, over 600 organizations formed the Anti-Phishing Working Group [3] focused on eliminating identity theft due to phishing.

Due to the rapid growth in the impact and number of phishing attacks, there is a considerable research effort going on both in academy and industry for developing robust and easy to use defense systems. Most of the currently available defense systems against phishing either limit the access of the user or display warning messages when they detect suspicious activities. Examples of such systems include e-mail spam filtering or browser plug-ins specially designed for monitoring user's transactions, e.g. SpoofGuard [4], Netcraft [5] or Ebay [6] toolbar. Another approach focuses directly on mitigating man-in-the-middle phishing attacks through a multi-factor authentication scheme [7]. We will briefly review these existing approaches in Section 2.2.

1.1 Our Approach

In this paper, we propose a defense system, ViWiD, that mitigates phishing attacks through an integrity check mechanism built on visible watermarking techniques. This mechanism is based on asking the user to check the validity of the visible watermark message on the logo images of the web pages. We propose two types of watermark messages: The first type is the time only watermark when the company's web site embeds only the current date and time of the user's time zone into the logo image. Recall that IP address can be used to determine the time zone of the user machine. An example of this type of watermarked logo can be seen in Figure 2[1](a). The second type of watermark message includes a secret shared between the user and the company together with the time stamp, as shown in Figure 2(b). The logo images with this shared secret watermark message can be displayed either after the user logs in, or through the usage of cookies. Since this watermarked logo displays a secret shared only between the user and the genuine company, the appearance of such information on the logo is enough for the user to confirm the genuineness of the web site.

The integrity checking system is designed to include a shared secret between the company and the user in order to prevent the phisher from performing the current "one-size-fits-all" attack. This means that even if the phisher is successful in removing the watermark, he can not insert back the expected watermark without knowing the shared secret between the company and the user.

The reasons for following this particular approach are as follows. First, phishing is primarily a social engineering attack which involves the active participation of the users to succeed. Thus, the approach towards mitigating such attacks must

[1] There is a quality loss in the displayed images through out the paper due to the conversion from Graphics Interchange Format (GIF) to Post Script (PS) format

also include the co-operation of the users to some extent. Indeed even today, the company web sites advise the users to follow well-known safety measures such as checking the padlock at the bottom of the screen and the 'https' sign in the URL, both of which signify a SSL connection. But, most of the victims of phishing attacks today are naive users who are not tech savvy enough to check the certificates or security sign. Also, the presence of a SSL connection by itself does not confirm the true identity of the web site. Any site, even a spoofed site, can establish a SSL connection. Communicating to naive users the true identity of the web site is a challenging problem. Hence, we propose the use of a shared secret which the user chooses himself when he registers with the original site. This shared secret can be easily recalled and recognized by the user. Using this secret, the company authenticates itself to the user. In the remaining of this paper, we will refer to this secret as a *mnemonic*.

Second, we chose *the web site logo* as a carrier for the watermark message, since the user always expects to see a logo on a web page. Besides this, the phisher always has to re-use the web site logo when he imitates the pages of the original web site. Since the original logos are always watermarked in our approach, it is not trivial for the phisher to remove them and insert his own watermarks. Even if the phisher is able to remove the watermark, he will not be able to insert the mnemonic for each user. More details about the proposed framework are presented in Section 3.

The rest of the paper is structured as follows. Next section provides a brief introduction to the anatomy of phishing attacks, state-of-the-art defense systems against phishing and summarizes the visible watermarking technique we use. We introduce our experimental set up and results in Section 4 and discuss possible attack models in Section 5. We conclude in Section 6.

2 Background

2.1 Phishing Attack Overview

In a typical phishing attack, a person receives an email apparently sent by an organization that the person interacted with before, and with which he has possibly built a trust relationship (e.g., his bank or a major retail on-line store). The email usually projects a sense of urgency, and asks the client to click on a link that, instead of linking to the real web page of the organization, will link to a fake web page that is subsequently used to collect personal and financial information. There are two victims in phishing attacks: the customer being tricked into giving away personal information and thus allowing the attacker to steal its the identity, and the company that the phisher is posing as, which will suffer both financial loss and reputation damage due to the attack.

Unauthenticated E-mail The major mechanism to start the attack is using forged e-mails. The phisher can forge e-mails by faking the source information displayed on the e-mail programs. Moreover, phishers can forge the content of the e-mail by getting a template of the style of legitimate e-mails when they subscribe to the company. The attack has a great impact because e-mail is the main communication channel for the online services. The subscribers or customers are expected to follow their transactions and receive confirmations via e-mails.

User Actions Phishing requires human interaction as like many of other online attacks do. However, unlike other attacks (worms or viruses spreading via e-mail) where one click is enough to trigger the attack, phishing requires active participation of the user at several steps, including providing personal information.

Deceptive View The core of the phishing attack lies in the ability of the phisher to create a web page looking very similar to a web page of the legitimate organizations by simply copying the logos, and using a style and structure similar to those on the legitimate page. In other words, the information displayed on web pages is not tied to its creator or owner in a way that removing that tie, will deteriorate the data beyond repair. In addition, many browsers are modifiable on the client side, allowing a phisher to remove buttons, not to display certain information, or to mislead the user by playing with the graphics.

A major challenge in addressing phishing attacks lies in designing mechanisms that are able to tie the data displayed on a web page (or related with a web page) to its legitimate owner. This is a difficult task because of the nature of the information displayed, its heterogeneous nature, and the dynamic characteristic of web pages.

2.2 Previous Approaches to Prevent Web-Based Identity Theft

Secure Email Many forms of phishing attacks can be prevented by the use of secure email tools such as Privacy Enhanced Mail (PEM), Secure Multipurpose Internet Mail Extension (S/MIME) and Pretty Good Privacy (PGP). However, to this date, secure email is not widely used over the Internet, because of scalability, trust, and difficulty to deploy it. A good discussion of certificate based security is provided in [8] by Ellison and Schneier.

Client-side Defense One direction in addressing the phishing attack was to provide the client with more accurate information about the web sites that he accesses. Various tools empowering clients with more information have been designed to mitigate phishing attacks. One such tool is SpoofGuard [4] which computes a spoof index and warns the user if the index exceeds a safety level selected by the user. The computation of the index uses domain name, url, link and image checks to evaluate the likelihood that a particular page is a spoof attack. One component of SpoofGuard maintains a database of hash of logo images and corresponding domain names. Later on a web page when the hash of the logo image matches a hash in the database, the current url is compared with the expected domain name, if these do not match the user is warned.

Netcraft, [5] also has released an anti-phishing toolbar that provides information about the web sites that are visited by a client such as the country hosting the sites and enforces the display of browser navigational controls (toolbar and address bar) in all windows.

Herzberg and Gbara [9] proposed establishing, within the browser window, a trusted credentials area (TCA). It is the browser that protects the TCA by preventing its occlusion. The scheme has its costs (it requires logo certification, logo certificate authorities, etc), but tolerates more naive users.

Cryptography-based Defense TriCipher, Inc. very recently introduced TriCipher Armored Credential System (TACS) against man-in-the-middle phishing attacks [7]. TACS works when the SSL client authentication is turned on. This means that the SSL protocol will have three steps: authenticate the web server to the client browser, set up encrypted communications and authenticate the end user to the web server. Common usage of SSL consists only of the first two steps. TACS uses two different types of credentials. The first one is called *double armored* credentials, and requires the users to install the TriCipher ID protection tool on their machine. The tool automatically pops up when the user goes to a page that is protected by SSL and encrypts (signs) the password using a key stored in the Trusted Platform Module or Windows® Key Store. Then the TACS appliance at the web server side authenticates the user. The second type of credentials is called *triple armored* credentials which uses, besides the user password and the key stored on the user's machine, a smart card or a USB memory stick to store a key or a biometric. The user's password is signed both with the key on the user's machine and another key stored elsewhere. The *triple armored* credential system raises the bar for the phisher because even if he is able to steal the key on the user's machine, he also has to steal the key stored on an outside system.

Shared Secret Schemes More recently two new authentication schemes, similar in nature to our system, have been brought to our attention. PassMark Security, Inc.'s [10] *2-Way Authentication Tool* helps the users identify known servers. In this scheme, the user provides the server with a shared secret, an image or a text phrase, in addition to his regular password. The server presents the user with this image, and the user is asked to recognize it before entering his password and authenticating himself to the server. Passmark images are randomly assigned to users from a pool of over 50,000 images and later the users can change their Passmarks, like they change their passwords, by selecting new images from the pool or by uploading an image of their choice.

In a very recent paper, Dhamija and Tygar proposed using Dynamic Security Skins [11] as a defense against phishing. Their system is based on having a *Trusted Window* in the browser and using the Secure Remote Password Protocol (SRP) [12] for authentication. Spoofing of trusted window is prevented by providing an image which is a shared-secret between the user and his browser. This window is dedicated to username and password entry. SRP is a *verifier-based protocol*. SRP provides the functionality for the server and user to authenticate each other over an un-trusted network by independently generating a session key based on a verifier. User sends the verifier to the server only once when he is registering. In Dynamic Security Skins, this verifier is used by the browser and the server to generate *a visual hash* that is displayed in the background of the trusted window and in the server's web site. To authenticate the server, the user needs to visually compare the two images to check if they match.

2.3 Limitations of Previous Approaches

Even though the client-side defense tools raise the bar for the attackers, they do not provide a complete solution. Many checks and enforcements used by the client-based defense tools can be fooled by attackers having a reasonable understanding of web site construction [4]. For example, the image check system of SpoofGuard can be fooled by a mosaic attack where the attacker partitions the logo image into pieces, but displays them in appropriate order so that the user thinks that he is looking at a legitimate logo.

Moreover, any "client side only" defense mechanism will suffer from false positives. Too many warnings will interfere with the user's browsing experience and the user will simply turn off the protection mechanism in such cases.

In addition to the above limitations, the "client side only" schemes leave all of the defensive actions and computational costs up to the user's machine, even though the companies have larger computing power at their disposal and can do more to mitigate the risks. Moreover it is the companies who create the content (logo, style etc) that the attackers seek to imitate and/or misuse. Therefore, we believe that companies can play a larger role in the overall defense strategy to mitigate phishing attacks.

On the other hand, cryptography based tools require the user to download a tool on every machine he uses to access his online accounts, and/or the user is required to carry another medium e.g. a smart card or USB memory stick with him when he wishes to access his accounts. One other limitation for TACS [7] is that it is designed to work only for man-in-the-middle phishing attacks. When the phisher directs the users to his web page which might have a SSL connection but without the client authentication module turned on, the TriCipher ID protection tool will not pop up and sign the password.

The shared secret schemes introduced in Section 2.2 are similar to our approach in the sense that they focus on how a legitimate server can authenticate itself to the user. However, our approach and these two approaches diverge on the generation and presentation of the shared-secret. The main drawback of PassMark approach is that the shared secret is not bound to a particular location on the original web page. This makes the scheme less user-friendly as across different service providers, the users will have to look at different places to find their shared secret on the web page. On the other side, Dynamic Security Skins [11] scheme suffers from asking the users to dedicate part of their browser window to the Trusted Window. Besides, this scheme trusts the client's browser on vital security processes such as storing the verifier and generating the visual hash.

Overall, a complete solution for defense against phishing, must address all three causes that allow a phishing attack to be possible: unauthenticated e-mail, user actions and deceptive view. Thus, a complete solution should include mechanisms that can analyze what the user sees, analyze the e-mail and web page content, and provide integrity checks for these components. In addition, such a system should be easy to use and deploy.

2.4 Visible Watermarking Overview

Visible watermarking is the insertion of a visible pattern or image into a cover image [13]. A useful visible watermarking technique should meet the following requirements: preserving the perceptibility of the cover image, providing reasonable visibility of the watermark pattern and robustness [14]. Huang and Wu summarize the insertion of a visible pattern into the cover image as:

$$I' = K_1 \cdot I + K_2 \cdot W \tag{1}$$

$$D(E_I(I'), E_I(I)) < Threshold_I \tag{2}$$

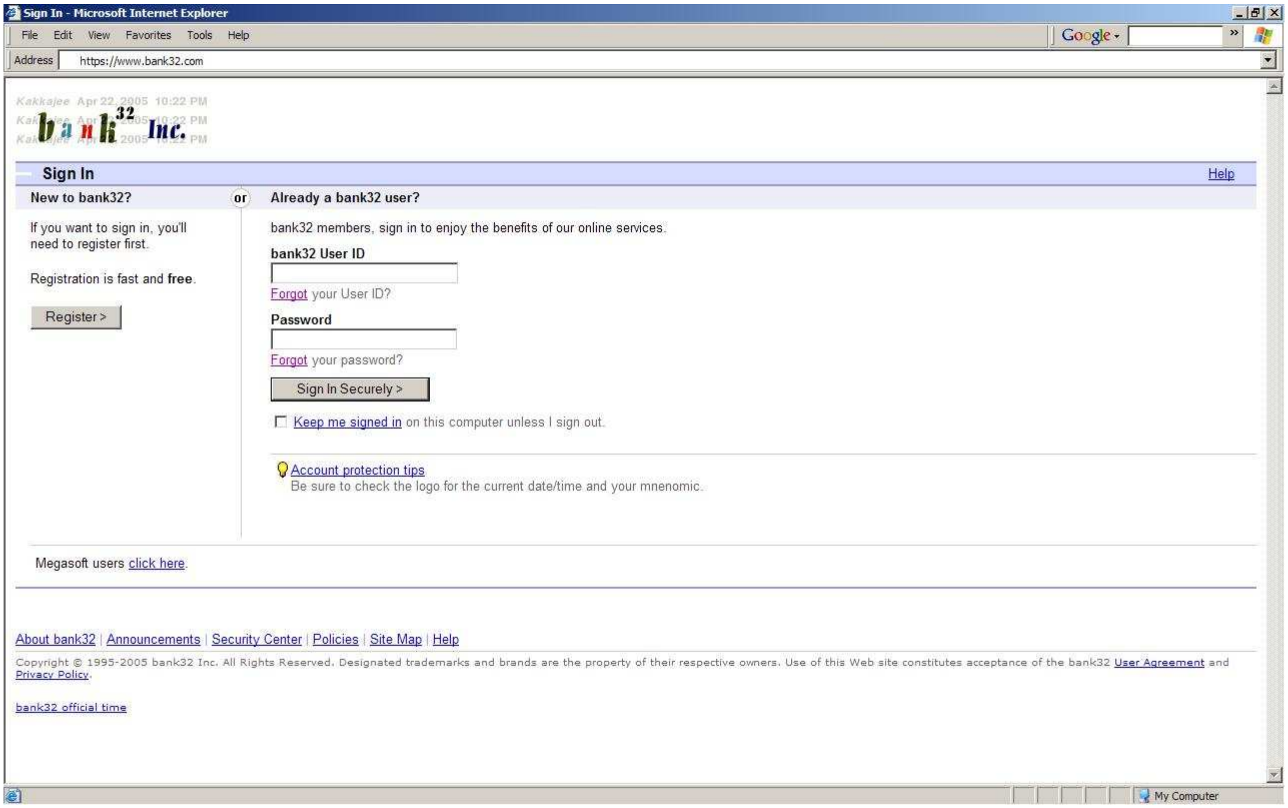
$$D(E_W(I'), E_W(I)) < Threshold_W \tag{3}$$

In Equation 1, I represents the cover image, W represents the watermark image and I' represents the watermarked image. Equation 2 represents the boundary on the distortion of the perceptibility of the cover image, while Equation 3 represents the boundary on the distortion of the visibility of the watermark patterns. D is a distance function measuring the perceptible difference of its two entries. E_I is a image feature extraction function for the cover and watermarked images. E_W is a separate image feature extraction function for the watermark pattern. $Threshold_I$ and $Threshold_W$ represent the largest allowable distortion on perceptibility of the cover image and on the visibility of the watermark pattern respectively.

In ViWiD, we use visible watermarking in order to provide the users with visibly watermarked logo images and the visible watermark pattern is generated dynamically depending on a shared secret between the user and the company.

3 Proposed Approach

The content of the e-mail and the spoofed page are the means through which the “social engineering” aspect of phishing is carried out. The phisher tricks the user into submitting sensitive information by using the content and the style stolen from the legitimate company. A good defense mechanism must require an integrity check method that “travels with the content” when it is used or misused. One way to achieve this is digital watermarking. Our approach watermarks the content on the legitimate web page in a way that provides an integrity check. We use the logo images as the watermark carriers, based on the observed fact that nearly all phishing attacks re-use the logo images.



New to bank32?

If you want to sign in, you'll need to register first.

Registration is fast and **free**.

Already a bank32 user?

bank32 members, sign in to enjoy the benefits of our online services.

bank32 User ID

Forgot your User ID?

Password

Forgot your password?

□ Keep me signed in on this computer unless I sign out.

Account protection tips
Be sure to check the logo for the current date/time and your mnenomic.

Megasoft users click here.

About bank32 | Announcements | Security Center | Policies | Site Map | Help

Copyright © 1995-2005 bank32 Inc. All Rights Reserved. Designated trademarks and brands are the property of their respective owners. Use of this Web site constitutes acceptance of the bank32 User Agreement and Privacy Policy.

bank32 official time

Fig. 1. A generic login page with a watermarked logo image, scaled to half of its original size for space requirements.

3.1 Design Goals and Motivation

The user can be tricked into a phishing attack, only if the phishing e-mail is imitating a company with which the user has previously established a trust relation. All companies, targeted by phishing attacks, have large numbers of users using their online services. Many of the users use several varieties of browsers and more than one computer to access their account online. A key-based watermark detection system requires the keys for detection and extraction to be distributed to all the users. We avoid the key distribution problem by using a visible watermark, with a human involved in the detection process. This way we also give the user an active role in the defense against a social engineering attack.

We seek to thwart the "one size fits all" attacks by designing the visible watermark message such that it is unique and varies with time. ViWiD embeds a local time stamp which is updated periodically and a mnemonic selected by the user while the online account established. The rationale for using the time stamps is that phishing sites are usually up for 6 to 7 days [3], and unless the phishers are able to remove the watermark, their stolen logo cannot display a fresh time to all the intended victims. Also, this system should never ask for the user's mnemonic after the online account is established in order to avoid the possibility of revealing the mnemonic even if the user mistakenly enters his login and password to a spoofed site.

3.2 Framework Description

On the publicly available web pages, the logo images display the date and time of the day as a visible watermark. An example is shown in Figure 2 (a). In these logo images, date and time are periodically updated to show the current time according to the user's time zone. The user will be trained to expect to see the current date and time as a visible watermark on the publicly available web pages.

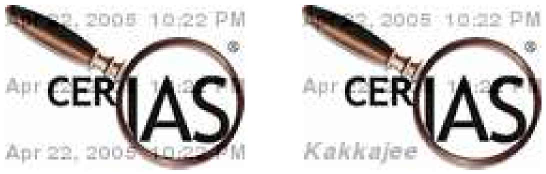
When the user establishes an account with the legitimate company, he is asked to select a mnemonic. We assume that there is a secure connection between the web server and the client side at that time to prevent the disclosure of the mnemonic to eavesdroppers. When cookies are enabled at the user's machine, the web site can use it to recognize the user the next time he is visiting the site. Using the cookie information, the web site knows which mnemonic to embed as a watermark in the logo images without authenticating the user. On the other hand, if cookies are disabled, then the mnemonic can only be added to the visible watermark after the user logs into the established account. This is a less satisfactory form of protection, as the alarm comes after the user has given his login and password. An example of a logo image carrying both the time stamp and the mnemonic is shown in Figure 2 (b).

In order to make the user expect these watermarks, the companies need to display messages that remind the user to verify the validity of the watermark displayed on the logo images. An example login page can be seen in Figure 1.

4 Experimental Set up and Results

We collected logo images from randomly selected web pages of 60 Fortune 500 companies and the Center for Education and Research in Information Assurance and Security (CERIAS). All of these logo images were colored Graphics Interchange Format (GIF) images. GIF is the preferred format for displaying logos on web pages because GIF images are 8-bit palette based images, hence their sizes are small. In our experiments, we tested the effectiveness of several visible watermarking algorithms on 61 logo images. The size of these logo images ranges from (18x18) to (760x50).

Even though there is a vast amount of literature on invisible image watermarking techniques, there have been relatively fewer visible image watermarking schemes developed to date [14]. We tested several different visible watermarking techniques on our logo images database. Visibly watermarking color logo images brings many challenges compared to watermarking gray scale images or JPEG images. The main challenge is to maintain the aesthetics of the watermarked logo so as to not to damage its marketing purpose yet be able to insert a robust and readable watermark into it. Moreover, visible watermarking on the logo images is rather less robust because these logo images have large uniform areas and very few objects in them. Besides these the time and memory requirements of the watermarking operation should be very low in order for the web server to be able to dynamically update the time stamp on the logo images frequently. We used the following two techniques in order to verify the applicability. In all these tests, we used a watermark image that is the same size as the cover image.



(a) (b)

Fig. 2. Logo images watermarked with $ImageMagick^{TM}$: (a) time only watermark (b) watermark with both time and mnemonic, in this image the mnemonic is *Kakkajee*.

- $ImageMagick^{TM}$**'s embedded watermarking module [15]** $ImageMagick^{TM}$ is a free software suite for the creation, modification and display of bitmap images. $ImageMagick^{TM}$ version 6.2.0 watermarking scheme updates brightness component of HSB color space of every pixel in the cover image using the following equations to embed the watermark:

  $$B'_{i,j} = B_{i,j} + \frac{(p \cdot offset^{w}_{i,j})}{midpoint} \tag{4}$$

  where $B'_{i,j}$ is the brightness of the watermarked image pixels, and $B_{i,j}$ is the brightness of the cover image pixels.

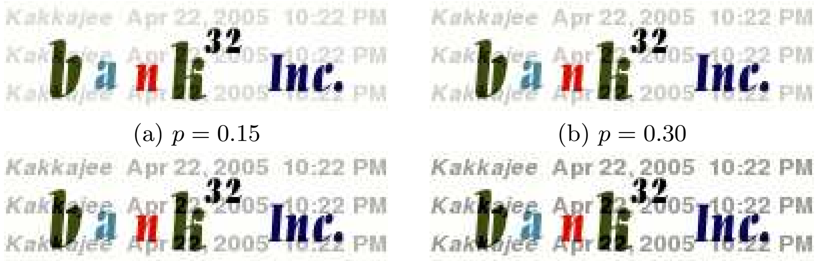
  $$offset^{w}_{i,j} = I^{w}_{i,j} - midpoint \tag{5}$$

  where $I^{w}_{i,j}$ is the intensity of the watermark image pixels.

  $$midpoint = \frac{maxRGB}{2} \tag{6}$$

  $maxRGB$ is the maximum value of a quantum, where a quantum is one of the red, green, or blue elements of a pixel in the RGB color space. In our experiments, $ImageMagick^{TM}$ was compiled with 16 bits in a quantum, thus giving $maxRGB$ equal to 65x535.
  p is a user selected parameter for the percentage brightness of the watermark pixel. An example of this embedding with $p = 0.3$ can be seen in Figure 2.

Hue and saturation of the cover image are not affected in the watermark embedding process. The value of the p parameter controls the visibility of the watermark. Figure 3 shows and example of the watermark embedding where the same watermark is embedded with varying p.
In order to preserve the aesthetics of the cover logo image, we used RGB (midpoint, midpoint, midpoint) as the background color in our watermark images. This is because, with these RGB values, the corresponding $offset^w$ values, in Equation 5, become 0.

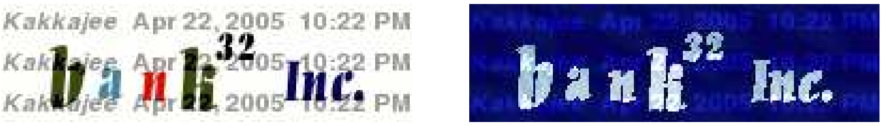


(a) $p = 0.15$ (b) $p = 0.30$

(c) $p = 0.30$ (d) $p = 0.40$

Fig. 3. Logo images watermarked with $ImageMagick^{TM}$ using various p values

We have observed that the background color, the text geometry on the watermark image and parameter p have to be adjusted according to the cover image properties in order to reach an acceptable level of watermarked image quality. Figure 4 shows examples of (a) a light and (b) a dark background logo images watermarked. In both Figure 4 (a) and (b) background of watermark image is RGB (midpoint, midpoint, midpoint) and $p = 0.40$. The color of the text of watermark image is black in Figure 4 (a), and white in Figure 4 (b).



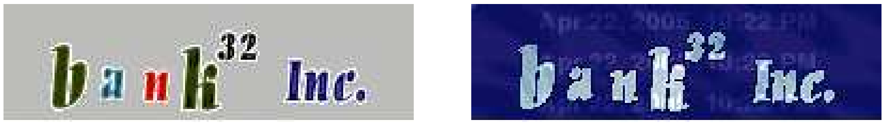
(a) (b)

Fig. 4. Logo images watermarked with $ImageMagick^{TM}$ parameter $p = 0.40$ (a) a white background and (b) a dark background

- **Mohanty et al.'s approach [16]** In their visible watermarking scheme, the modification of the gray values of the host image is based on its local as well as global statistics.

$$I'_n = \alpha_n \cdot I_n + \beta_n \cdot I^w_n \tag{7}$$

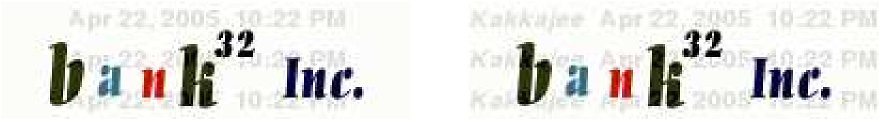
where I'_n is the intensity of the n_{th} block of the watermarked image. I_n and I^w_n are the corresponding intensity values of the cover and watermark images respectively. α_n and β_n are the scaling and embedding factors depending on the mean and the variance of each block, and the image mean gray value. In [16], it is stated that for color images the watermark should be put in the Y component (luminance). However, when this approach is applied on logo images with white background, even a small change in the luminosity of the background will disturb the aesthetics of the logo image. An example of this phenomenon can be seen in Figure 5 (a). On the other hand, logo images with dark background gave better results, see Figure 5 (b) for an example.



(a) (b)

Fig. 5. Logo images watermarked with Mohanty et al.'s watermarking algorithm (a) a white background and (b) a dark background

However, we observed that the K component of the $CMYK$ colormap can also be used to insert the watermark into logo images. This modified approach gave us better results on logo images with white background, see Figure 6.



(a) (b)

Fig. 6. Logo images watermarked with modified version of Mohanty et al.'s watermarking algorithm (a) Time only watermarked logo image (b) Watermarked logo image with Time and Mnemonic

We are not able to provide samples from the watermarked version of the logo images we collected from Fortune 500 companies' web pages due to copyright issues. In addition, there is a quality loss in the displayed images through out the paper due to the conversion from GIF to Post Script (PS) format. In order to provide GIF versions of the watermarked logo images and a controlled access to these logo images, we have created a demo page which can be reached at http://projects.cerias.purdue.edu/viwid/.

5 Security Analysis and Discussion

A phisher can try to break the above system through the following three attacks. First attack is to insert a valid watermark message after removing the existing watermark from the logo image. The second attack is to recreate the logo image from scratch and later insert a valid watermark message. The third attack is to perform a man-in-the-middle attack. We explain below why these attacks are not easy for an attacker to carry out.

Success of the first attack depends on the robustness of the underlying visible watermarking algorithm and on the success of the phisher at generating the valid watermark messages for the targeted users.

Huang and Wu , in [14], show successful attacks on well known visible watermarking systems [16, 17] with the help of human intervention. Huang and Wu's system requires the shapes of the watermark patterns to be marked manually. Results in [14] show that the image inpainting techniques are very effective in removing simple watermark patterns composed of thin lines or symbols. For more sophisticated watermark patterns such as thick lines or bold faced and multi-textured text, Huang and Wu propose an improved scheme where thick watermarked areas are classified into edges and flat areas. Later flat watermarked areas are recovered by refilling them with unaltered flat neighbours. Edged watermarked areas are recovered by approximated prediction based on adaptation information of nearby pixels.

However, in ViWiD, even if the attacker is able to remove the watermark successfully from the watermarked image, he can not insert a completely valid watermark message. The valid watermark message consists of the date and local time of the day for the user's time zone, and the user's mnemonic. The mnemonic is unique for every user and the attacker does not have access to any user's mnemonic. If he can have such access, his attack ceases to be a "one-size-fits-all", and thus we have succeeded in increasing the attacker's cost.

The second attack, which requires recreating the logo image from scratch, can also be thwarted by the fact that the attacker is unable to generate the valid watermark message for every user.

The man-in-the-middle attack is one of the most successful ways of gaining control of customer information [18]. However, besides directing the user to his machine through social engineering, it is difficult for the phisher to be successful in this attack. He has to either manipulate the DNS or proxy data on the user's machine, or locate the attacking machine on the real company's web server's network segment or on the route to the real company's web server. Even if the phisher performs a man-in-the-middle attack in order to bring a fresh logo every time a user requests the phisher's web page, the web site would only provide the logo specifically watermarked for the time zone that is assigned to the attacker's

IP address. In such a case the attacker would need to have available as many man-in-the-middle's as the number of time zones he wants to attack.

6 Concluding Remarks

We have presented a defense system, ViWiD, that mitigates phishing attacks through integrity checking of web site logos using visible watermarking techniques. The valid watermark message consists of the date and local time of the day for the user's time zone, and the user's mnemonic. The watermark message is designed to be unique for every user and carries a shared secret between the company and the user in order to thwart the "one size fits all" attacks.

Unlike the other systems proposed for preventing phishing attacks, ViWiD performs all of the computation on the company's web server and does not require installation of any tool or storage of any data, such as keys or history logs, on the user's machine. ViWiD also involves the user in the integrity checking process, which makes it harder for the phisher to engineer an attack, since the integrity checking mechanism is not fully automated.

One of the pre-requisites of the proposed scheme is that it requires the users to be trained to expect a valid message to be displayed on the logo images when they perform sensitive transactions. Users are also provided the opportunity to adjust the parameters of the watermark and logo image according to their reading needs and appeal. For example, a user might select a larger font size for the embedded watermark message, or he can as well select a larger logo image.

As part of future work, we plan to perform a large scale user study for validating the effectiveness of our approach. In addition to that, the robustness of the watermarking techniques can be improved by using high quality logo images in JPEG format or by spreading the message over all images in a web page.

7 Acknowledgments

The authors would like to thank Chris Baker, Umut Topkara, Eugene Lin, Prathima Rao, Ardalan Kangarlou-Haghigh and Saraju Mohanty for their helpful comments.

References

1. S. M. Martinez, "Identity theft and cyber crime," September 2004. Federal Bureau of Investigation, http://www.fbi.gov/congress/congress.htm.
2. "Symantec internet security threat report highlights rise in threats to confidential information," *http://www.symantec.com/press/2005/n050321.html*.
3. "The Anti-Phishing working group," *http://www.antiphishing.org*.
4. N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, "Client-side defense against web-based identity theft," *Proceedings of the Network and Distributed System Security Symposium*, 2004.

5. "Netcraft," *http://www.netcraft.com.*
6. "ebay: buyer tools: toolbar," *http://pages.ebay.com/ebay_toolbar/.*
7. "Preventing man in the middle phishing attacks with multi-factor authentication," *http://www.tricipher.com/solutions/phishing.html.*
8. C. Ellison and B. Schneier, "Inside risks: Risks of PKI: secure email," *Communications of the ACM*, vol. 43, no. 1, p. 160, 2000.
9. A. Herzberg and A. Gbara, "Trustbar: Protecting (even nave) web users from spoofing and phishing attacks," *Cryptology ePrint Archive, Report 2004/155*, 2004.
10. P. Security, "Protecting your customers from phishing attacks - an introduction to passmarks," *http://www.passmarksecurity.com/.*
11. R. Dhamija and J. Tygar, "The battle against phishing: Dynamic security skins," *Symposium on Usable Provacy and Security (SOUPS)*, July, 2005.
12. T. Wu., "The secure remote password protocol," *In Internet Society Network and Distributed Systems Security Symposium (NDSS)*, Mar 1998, pp. 97–111.
13. N. Memon and P. W. Wong, "Protecting digital media content," *Commun. ACM*, vol. 41, no. 7, pp. 35–43, 1998.
14. C.-H. Huang and J.-L. Wu, "Attacking visible watermarking schemes," *IEEE Transactions on Multimedia*, vol. 6, no. 1, February, 2004.
15. "Imagemagick studio llc," *http://www.imagemagick.org.*
16. S. P. Mohanty, K. R. Ramakrishnan, and M. Kankanhalli, "A dual watermarking technique for images," *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 2)*, 1999, ACM Press, pp. 49–51.
17. G. W. Braudaway, K. A. Magerlein, and F. C. Mintzer, "Protecting Publicly Available Images with a Visible Image Watermark," *Proceedings of the SPIE International Conference on Electronic Imaging*, vol. 2659, February 1-2, 1996, San Jose, CA.
18. G. Ollman, "The phishing guide," *http://www.ngssoftware.com/papers/NISR-WP-Phishing.pdf.*