Phishing is an emerging type of social engineering crime on the Web. Most phishers initiate attacks by sending emails to potential victims. These emails lure users to access fake websites, and induce them to expose sensitive and/or private information. Phishing has become a severe problem of internet security. We propose a phishing web page detection method using the EMD based visual similarity assessment. This approach works at the pixel level of web pages rather than at the text level, and can thus detect phishing web pages only if they are “visually similar” to the protected ones without considering the similarity of the source codes. Our experiments also show that our method can achieve satisfying classification precision and phishing recall, and the time efficiency of computation is acceptable for online use.
This document clearly describes the problem statement, requirements, issues that are handled, modules involved and also the assumptions with which the components are implemented. These are important particularly to ensure our information. We can save our information from various attack and we can make good decision for web pages whether it is true or false.

**Architecture:**

![Figure 1.1 General Architecture](image)

We propose an effective approach to detecting phishing web pages. We employ the Earth Mover’s Distance (EMD) to calculate the visual similarity of web pages. The most important reason that internet users could become phishing victims is that phishing web pages usually have high visual similarity with real web pages, such as visually similar block layouts, dominant colors, images, and fonts etc. We follow the anti-phishing strategy in to obtain suspected web pages. We parse URLs in suspected emails, obtain web pages of these URLs, and convert them into normalized images. We represent signatures of these images with features composed of the dominant color category and its corresponding centroid coordinate to calculate the visual similarity to protected web pages. The linear programming algorithm is used to calculate EMDs. EMD represent the dissimilarity of two signatures. An anti-phishing system can protect many web pages simultaneously. A threshold is calculated for each protected web page using supervised training. If the EMD based visual similarity of a web page exceeds the threshold of a protected web page, we classify the web page as phishing. Experiments on a large scale have been carried out. Our experiments showed high classification precision, high phishing recall, and satisfactory time performance. We also made a performance comparison to the method proposed by Liu et al in and the HTML based EMD method. Our experimental results showed that the proposed method is superior to the other two.

**EMD based Visual Assessment**

EMD is a method to evaluate the distance (dissimilarity) between two given signatures. A signature is a set of features and their corresponding weights. The method comes from a well known transportation problem. Suppose we have \( m \) producers, each with a weight representing the amount of product it has.

We denote producer set \( P \):\( \{(p_1,p_2), \ldots, p_m\} \)

and consumer set \( C \):\( \{(c_1, c_2), \ldots, c_n\} \)

the EMD based visual similarity of a web page exceeds a threshold, then we classify the web page as phishing.

Suppose we also have \( n \) customers, each with a weight indicating the amount of product he needs. We denote the consumer set \( C \):\( \{(c_1, c_2), \ldots, c_n\} \)

Producers want to transport their products to consumers. Suppose the distances of each pair of producer and consumer are known, and they are represented into a distance matrix \( D \), which is defined before calculating EMD.

It is represented as: \( D = [d_{ij}] \), where \( 1 \leq i \leq m \) and \( 1 \leq j \leq n \)

Producers all produce the same product and consumers all consume the same product. The transportation fee is proportional to both distance and product weight. The task is to find a flow matrix \( F \), which contains factors indicating the amount of product to be moved from one producer to one consumer.

\( F = [f_{ij}] \), where \( 1 \leq i \leq m \) and \( 1 \leq j \leq n \)

The transported product amount from \( P \) to \( C \) should be as much as possible and the total transportation fee should be minimized.

**Web page Preprocessing and Signature Generation**

We retrieve the suspected web pages and protected web pages from the web and generate their signatures. The preprocessing of web pages is a three step process:

- Obtain the image of the web page from the given URL;
- Perform normalization; and
- Represent the web page image as a web page visual signature (consists of color and coordinate features).

The process of displaying a web page in a web browser on the screen from HTML and accessory files (including pictures, Flash movies, ActiveX plug-ins, Java Applets, etc.) is the web page rendering. We use the GDI (graphic device interface) API provided by Microsoft Internet Explorer to get web page images and save them as jpeg files at original size. The images of the original sizes are transformed into images with normalized sizes (e.g. 100*100). The Lanczos algorithm is used to calculate the resized image because it has very strong anti-aliasing properties in Fourier domain, and it is also easy to compute in spatial domains. Lanczos algorithm can also generate sharp images, and intuitively, sharp images can provide clearer signature for images. Figure 1.2 shows examples of original web pages and resized images (to 100*100 and 10*10, respectively). www.bbb.org is an example of a square-like image, www.banktechnews.com is an example of a longer image, and www.bankofcyprus.com is an example of a wider image. Each of them is normalized into fixed-size square images. A signature of an image is a feature vector which can effectively represent the image. The Signature of an image used in our approach comprises features and their corresponding weights. A feature comprises a color and the centroid of its position distribution in the image.
The color of each pixel in the resized images is represented using the ARGB (alpha, red, green, and blue) scheme with 4 bytes (32 bits). A color can be represented with a 4-tuple \( < A, R, G, B > \). However, this is a huge color space, which includes \( 2^{32} = 4,294,967,296 \) colors. In practice, we use a degraded color space to represent the signature of an image. We define the Color Degrading Factor (CDF) to be the scale of each color component making a change. Thus, we have \( \frac{28}{\text{CDF}} \) colors in our degraded color space. A degraded color can be represented as \( < A-(A \mod \text{CDF}), B-(B \mod \text{CDF}), C-(C \mod \text{CDF}), D-(D \mod \text{CDF}) > \).

The centroid of each degraded color is calculated using:

\[
C_{dc} = \sum_{i=0}^{\text{Ndc}-1} \frac{f_{dc,i}}{\text{Ndc}},
\]

where \( C_{dc} \) is the centroid of degraded color \( dc \), \( f_{dc,i} \) is the coordinates of the \( i \)th pixel that has degraded color \( dc \), and \( \text{Ndc} \) is the total number of pixels that have degraded color \( dc \), i.e., the frequency of \( dc \). A feature \( F_{dc} \), which has degraded color \( dc \) can be represented with \( dc \) and \( C_{dc} \), \( F_{dc}=< dc, C_{dc} > \). The weight corresponding to this feature is the color’s frequency \( \text{Ndc} \). A complete signature \( S \) is represented as \( S=< S_1, S_2, \ldots, S_m > \), where \( S_1, S_2, \ldots, S_m \) has \( m \) features.

Computing Visual Similarity from EMD

We use EMD to calculate the similarity of two web pages based on their signatures as follows. The distance matrix \( D=[d_{ij}] \) ( \( 1 \leq i \leq m, 1 \leq j \leq n \) ) is defined in advance in a straightforward way: we first calculate the normalized Euclidian distance of the degraded ARGB colors, and then calculate the normalized Euclidian distance of centroids. The two distances are added up with weights \( p \) and \( q \) respectively to form the feature distance, where \( p+q=1 \). Suppose we have feature,

\[
\omega_i = < dc_i, C_{dc_i} >
\]

\[
dc_i=< dA_i, dR_i, dG_i, dB_i >
\]

Where

\[
\omega_j = < dc_j, C_{dc_j} >, \quad \text{where} \quad dc_j=< dA_j, dR_j, dG_j, dB_j >
\]

the maximum color distance \( MD = \|< \text{Max}A - 0, \text{Max}R - 0, \text{Max}G - 0, \text{Max}B - 0 >\| \) color, where \( \text{Max}A, \text{Max}R, \text{Max}G, \text{Max}B \) are the maximum numbers of the four components of ARGB respectively in the specified color space. The normalized color distance \( ND_{\text{color}} \) is defined as

\[
ND_{\text{color}}(dc_i, dc_j) = \frac{\sqrt{(dA_i - dA_j)^2 + (dR_i - dR_j)^2 + (dG_i - dG_j)^2 + (dB_i - dB_j)^2}}{MD_{\text{color}}}
\]

The normalized centroid distance \( ND_{\text{centroid}} \) is defined as

\[
ND_{\text{centroid}}(C_{dc_i}, C_{dc_j}) = \frac{||C_{dc_i} - C_{dc_j}||}{MD_{\text{centroid}}}
\]

The normalized feature distance between \( S_a, S_b \) is

\[
ND_{\text{feature}}(\omega_i, \omega_j) = p \cdot ND_{\text{color}}(dc_i, dc_j) + q \cdot ND_{\text{centroid}}(C_{dc_i}, C_{dc_j})
\]

Suppose we have signature \( s_a \) and signature \( s_b \), where \( s_a \) has \( m \) features and \( s_b \) has \( n \) features. The flow matrix \( F_{ab}=[f_{ij}] \) ( \( 1 \leq i \leq m, 1 \leq j \leq n \) ) can be calculated through linear programming, and the EMD, can be calculated as:

\[
EMD(S_a, S_b, D) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \cdot d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}
\]

If EMD value is 0 the two images are identical, and if EMD value is 1 the two images are completely different.

II. CONCLUSION AND FUTURE WORK

Phishing causes severe problems for Web security and privacy. In this dissertation, we address several advanced anti-phishing technologies systematically. We address the potential phishing attacks and propose to solve these phishing attacks with visual assessment approach. Phishers usually make web pages visually similar to real web pages to spoof users. We use Earth Mover’s Distance to evaluate the visual similarity of the suspected web pages to the protected ones. The suspected web pages which are similar to the protected ones will be reported as phishing. Our experiments...
show that this method performs quite well. This method may fail to work when phishers intend to make the phishing web pages different from the real ones.
Phishing attacks have severe negative impacts for the Web and are also are evolving criminal techniques. We address several advanced anti-phishing methods in this dissertation and their counter measures. However, we are not expecting that all phishing problems can be solved by these methods. Other phishing methods would be carried out in the future. Hence, new specific anti-phishing methods should be proposed.

REFERENCES
5. Anti-Phishing Group of City University of Hong Kong, http://antiphishing.cs.cityu.edu.hk
7. Ashok A., Jakobsson M., Detecting Obfuscation of malicious email by Virtual Rendering, manuscript in preparation, School of Informatics, Indiana University at Bloomington