

# **ECE 1776 – Project Midterm Report**

## **Overview of Project**

Our anti-phishing extension has been challenging up to this point. To help define our problem, we have analyzed many variations of the phishing problem. Our team has created a Firefox extension that queries the Google search engine using a variety of query terms. We are interested in using Google as a trusted online reputation system to validate the trustworthiness of web sites and to filter out phishing web sites versus valid web sites. We have observed that phishing web sites continue to evolve and deceive unsuspecting users into providing their user ID and passwords in a variety of new and interesting ways. This mid-term report will provide an update on the results we have obtained using our anti-phishing prototype extension, interesting problems and situations that we have come across in our work and outline the future plans of this project.

## **Current Status**

Our team has created an initial prototype of our Firefox extension that parses and then queries a given web page using the Google API. We have added more features in each development iteration so that the extension can now identify login pages of a web site by inspecting HTML form elements in the document object model (DOM) of a particular web page. Once the web page has been identified as a login page, the extension parses DOM elements such as the title, meta-tags and alternative image text. These values are then used as the keywords to query Google to get a list of reputable sites that appear in the top 20 list of search results. The extension then checks if there is a match between the current site and Google's search results. We have started to test our prototype using a variety of valid financial services and popular e-commerce web sites as well as suspected phishing web sites using a new phishing identification service called Phishtank [2].

## **Interesting Problems**

One difficult aspect to identify fake versus authentic web sites was choosing the correct or most relevant keywords to query with Google. Since most web sites are unique and inconsistent we had to look at many web sites to see which terms and HTML elements most commonly appear. We decided to focus on these sources of information in the web site to look for keywords:

1. Meta data - This HTML element allows authors to specify meta data information about a document in a variety of ways [1]. Meta data is aimed to help search engines index documents by defining properties such as "keywords" and "descriptions".
2. Title - Every HTML document must have a TITLE element in the HEAD section [1]. Authors should use the TITLE element to identify the contents of a document.
3. Alternate text - The web site authors are required to provide an image description with the ALT attribute, moreover they are required to specify meaningful alternate text [1]. For example, the alternate text for a company's logo should be ``.

The HTML meta data, title and ALT attributes are well defined and two of them are required; in practice, we have seen that most pages do not provide useful information or any information at all in these tags. The meta data tags provide keywords and descriptions of the site, but usually this information is verbose, thus we decided to select a subset of the information. Another point is that sometimes meta data is not always present. The title tag has resulted in a good source of keywords, most of the time both authentic and fake web sites provide useful information in the title. We have noted that company names usually appear in the title tag but there are cases where title tags do not provide any relevant keywords. For example, one web site has the title of, "Online Banking" [3]. Finally, if alternate text appears for one image, it will usually appear for every image in the web page. Instead of using the text for every available image, we decided to use the text that appears in the first three or four images, because in general the first images are related to the company's logo. In our analysis of web pages, we noted that most web sites have only one relevant Meta data, title or alternate text element which can be used for the keywords to query Google. Thus having an extension that queries these three elements will help to identify most web pages using the Google search engine. Also, we have processed pages that contain a plethora of information in these elements, in this case we have had to select just a few keywords to use in our Google search.

The prototype is able to query Google with the selected keywords that we extract from the web page. We decided not to compare the complete site URL with the returned Google URL since the search results will return the homepage of the web site and may not be the same URL as the login page. Therefore, we compare the domain name of the site with the domain names of the search results. If there is a match, we know that the specific web page resides in a reputable domain. Nevertheless, sometimes we were not able to identify authentic sites, because of a similar problem described above. In one example, the results from a Google query gave us the main domain name of the company,

but the login page we were inspecting was hosted on a sub-domain. For example, the authentic CIBC Online Banking login web site is located at *www.cibconline.cibc.com*, when we query Google with the keywords: *Online Banking CIBC CIBC Banking*. The top 20 results from Google return pages that are in the *www.cibc.com* domain [3]. Thus, our comparison between domain results was not successful, even when the two web sites were from the same company. In order to solve this problem, we decided to implement an algorithm that takes advantage of the domain name hierarchy, thus we are able to match a web site's sub-domain to the actual domain that appears in the Google search results.

### **Plans for the Remainder of Semester**

We plan to continue on improving our prototype through further testing and modifications to our extension. First, we plan to explore the use of additional information sources that can be used to identify fraudulent web sites, such as the Google page rank values and evaluating URLs that contain IP addresses. We also want to explore different options of the Google API that will help us fine tune our search algorithm such as using extended search techniques. We also plan to create a way to display a notice to the user that a suspected phishing site has been identified. Our current prototype only provides this information in the Firefox console logs. To ensure our prototype is identifying correct phishing web sites, we plan to continue testing our extension with valid and phishing web sites. These results will help us identify positive, false positive, and false negative phishing web sites. We will then use this information to help further refine our algorithm to identify phishing web sites. Lastly, we will continue to review the literature and news related to the phishing problem to help determine if we are approaching it from the right direction.

### **References**

[1] HTML 4.01 Specification, <<http://www.w3.org/TR/html401>>

[2] PhishTank. Out of the Net, into the Tank <<http://www.phishtank.com>>

[3] CIBC Online Banking <<https://www.cibconline.cibc.com>>