***Chapter 1***

***Introduction***

Internet has become another world itself in which you can get everything which is sufficient to live a life. You can watch movie, order food, earn money make friends and find your soul mate even. It has created huge revolution in our life. According to statics, almost 3 billion users have access of internet. Unfortunately, this world has become very vulnerable and unsafe. Internet banking, ecommerce, email services share significant amount of usage of this modern world. Often information and data being transmitted through the internet is very valuable and confidential. Poor security and some vulnerabilities have made easy to gain access of such confidential data for bad programmers.

Among all of these security attacks, Phishing attack is known for stealing private information. According to *Kaspersky Lab’s* database, 1 million number of phishing attacks has been increased in the first quarter of 2015 compare to previous quarter [1]. World is becoming online which has increased the number of websites so as phishing attacks. Phishing attacks indirectly affect many well know organization’s reputation. Many solutions were provided till now to detect phishing websites. None of those solutions provides good accuracy and performance when it comes real time safe browsing.

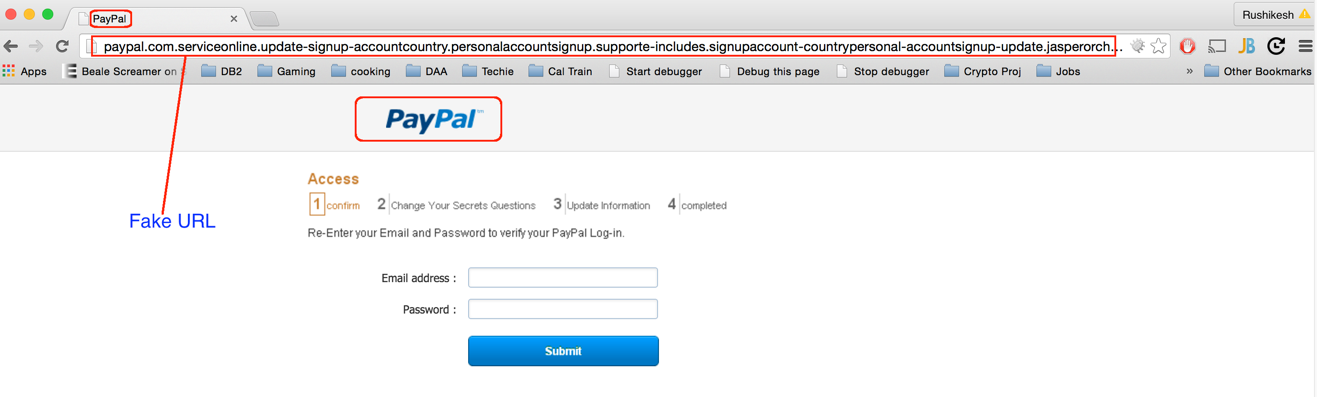
***1.1 What is phishing?***

Phishing can be defined as activity of collecting unauthorized and confidential data such as username, password, credit card detail, bank account details electronically. First time phishing activity was defined in detail in a paper and presentation delivered to the 1987 International HP Users Group, Interex [2]. First phishing attack was registered by AOHell hacking tool [3].

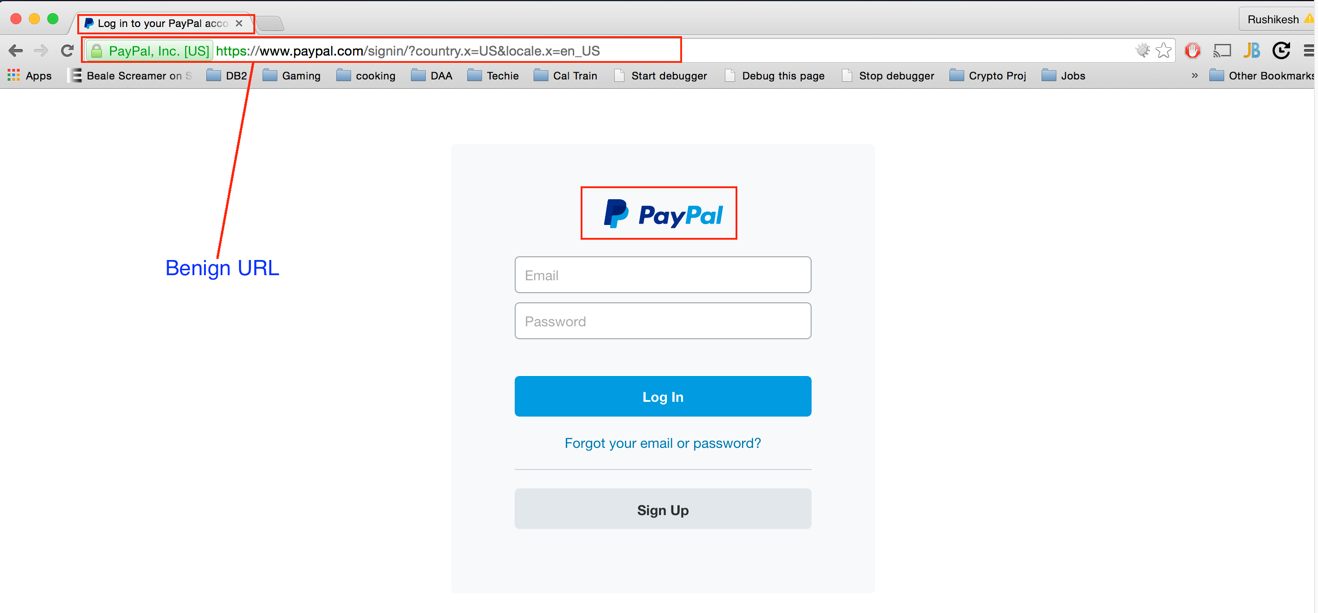
There are various types of phishing attacks. Phishing can be done in numerous ways. Following is the type of phishing attacks.

1. Link manipulation
2. Filter evasion
3. Website forgery
4. Covert redirect
5. Phone phishing

Our main focus in on “Link manipulation” type of phishing attack. In this attack URLs appear to belong to the valid organization. URLs are obfuscated very smartly that it is difficult sometimes to differentiate by human eye also. Let’s take an example of such website.



1. Paypal fake website



1. Paypal original website

As you can see in the first image this site claims to be genuine Paypal, a worldwide online payments system company. Second image is the real original website of Paypal. If you closely look then you easily see the difference between two websites. There are some visible difference between these two websites like logo of the company, favicon and secured certificate. Different types of techniques is used to redirect such fake websites. Sometimes users do not see this visible difference and becomes the victims of this phishing attacks.

Once the data is collected, different types of forgeries is done by hackers. This scam can be of Millions of money sometimes. Sometimes its all about private and confidential information of celebrity which can be leaked to spoil the image of the same person.

***1.2 Problems***

To address this Phishing problems lots of academic and business research has been done so far. These methods either uses blacklist methods or some type of heuristics or machine learning techniques. Blacklist method is more common and popular for all the browser. All browser contains some manually verified list of phishing websites. This technique contains fairly low positive rate. But when it comes to fresh phishing website, it does fail. It is not much effective to newly developed phishing websites.

Second approach is to adopt smart heuristics which is given some training data to train the heuristic. This heuristic collects different types of features of the websites and based on that it decides the authenticity of the website. So far no heuristic has defined which has better accuracy, performance and seamlessly browsing experience. In addition, scope of the detection of phishing website should not be limited. We referred to similar kind of one research paper which talks about all of these features. We tried to collect all the mentioned attributes and experimented with different machine learning classifiers. This method highly depends on approximate string matching algorithms and WHOIS server queries. In the development procedure we used only one third party services which is WHOIS server. Otherwise every algorithm run on native system. Our tests with this method shows some promising results. With the help of this attributes and approach we were able to achieve almost 97% accuracy and comparatively good performance with respect to time.

One more research paper has proposed to use image processing techniques. They tried to emphasize on the favicon of the website. They proved that significant difference is found when you compare the favicon of original website and spoofed website. Limitation of this approach is website must have favicon otherwise this method will fail.

***1.3 My Contribution***

My main role was go give practical shape to the idea which they talked in the research paper. Since everything was on the paper only and they did not talk about any implementation part, it required lot of efforts to make it practical.

There are couple of different way to make things happens. We split this entire addons in three different parts. First component resides on client side. Second component is web services which is kind of middle layer between third component, machine learning algorithms and addons.

We opted for making firefox addons. It has got pretty nice documentation to start with. Then there were couple of options available for machine learning libraries. Weka seemed quite distant options for us. It provides api for different languages.

We collected different samples of phishing websites from Phishtank[] which is famous organization for providing database of phishing websites. We used Alexa to get samples of benign websites. Total 1773 URLs collected for training data set. Out of 1773 URLs 829 URLs were phishing and 944 URLs were benign. We tried to apply different machine learning classifiers to get the best accuracy. As per our results, our filter achieves almost 97% accuracy to detect phishing websites.

***Chapter 2***

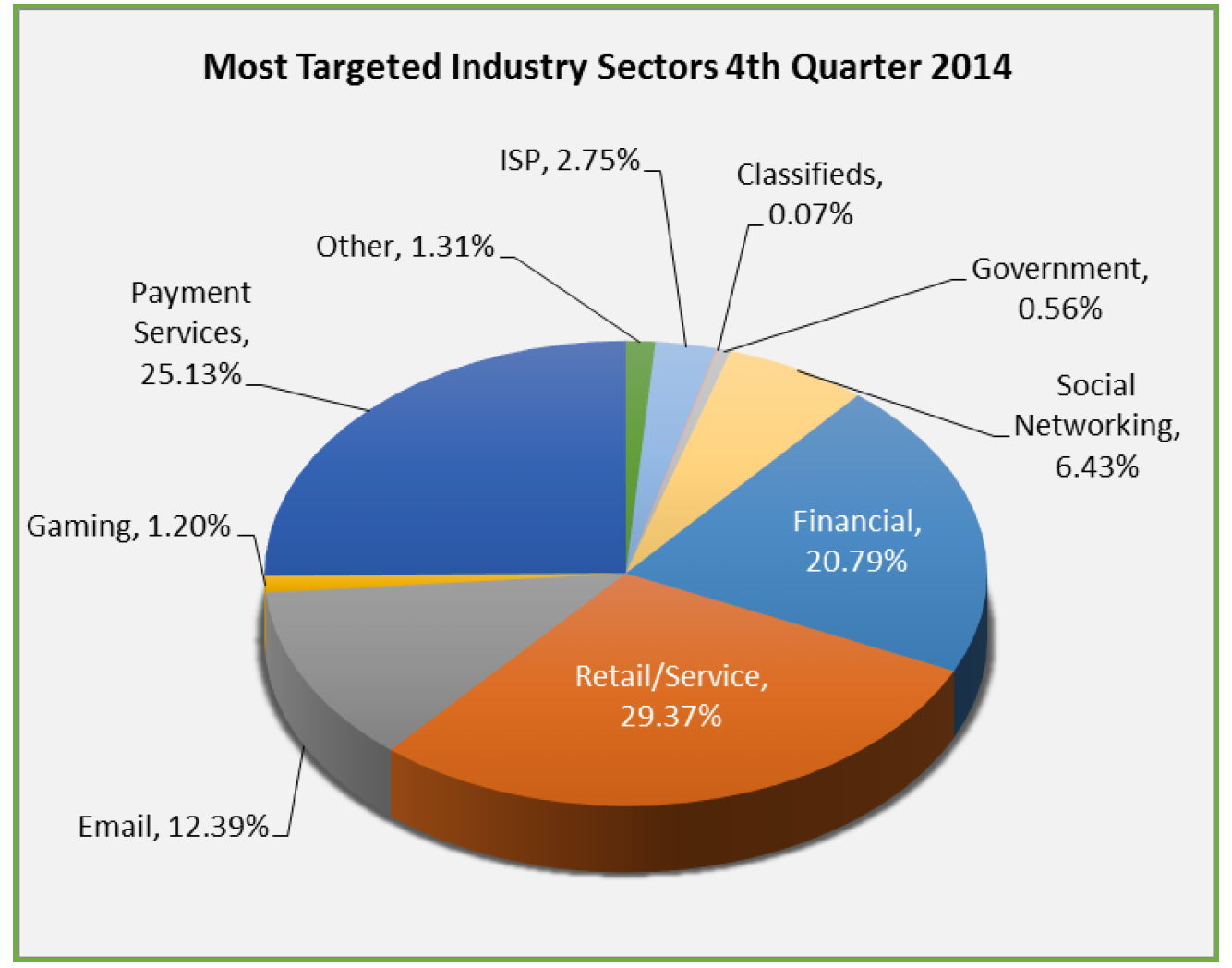
***Background***

***2.1 How big problem is***

The rate of private and sensitive information transmitted through internet has increased exponentially from last couple of years. People are now getting used to save all of their information on the cloud which has inspired phishing a lot. As you know the main purpose of the phishing is to steal any kind of information. So, the usage of cloud will increase, lot of data will be there to steal. The purpose of the phishing attack could be anything like to hack personal account, to spy on someone’s personal life etc.

The “Anti Phishing Group” is an international association focused on unifying the global response to cybercrime. This group provides a platform to discuss cybercrime issues majorly related to “Phishing”, to discuss potential technology solution to this problem, to access data related to cybercrime forensics. This group publishes a quarterly report. As per their quarter 4 of 2014 report, “Retail/Service” is the most popular industry sector for hackers. As you can see the chart second and third most popular industries are “Payment Services” and “Finance”. If we sum up the both sectors, then almost 45 percent of the attacks are in this industries which includes billions of money in transactions. We can easily say the reason is money. As per rough estimate millions of money is lost because of this attack.

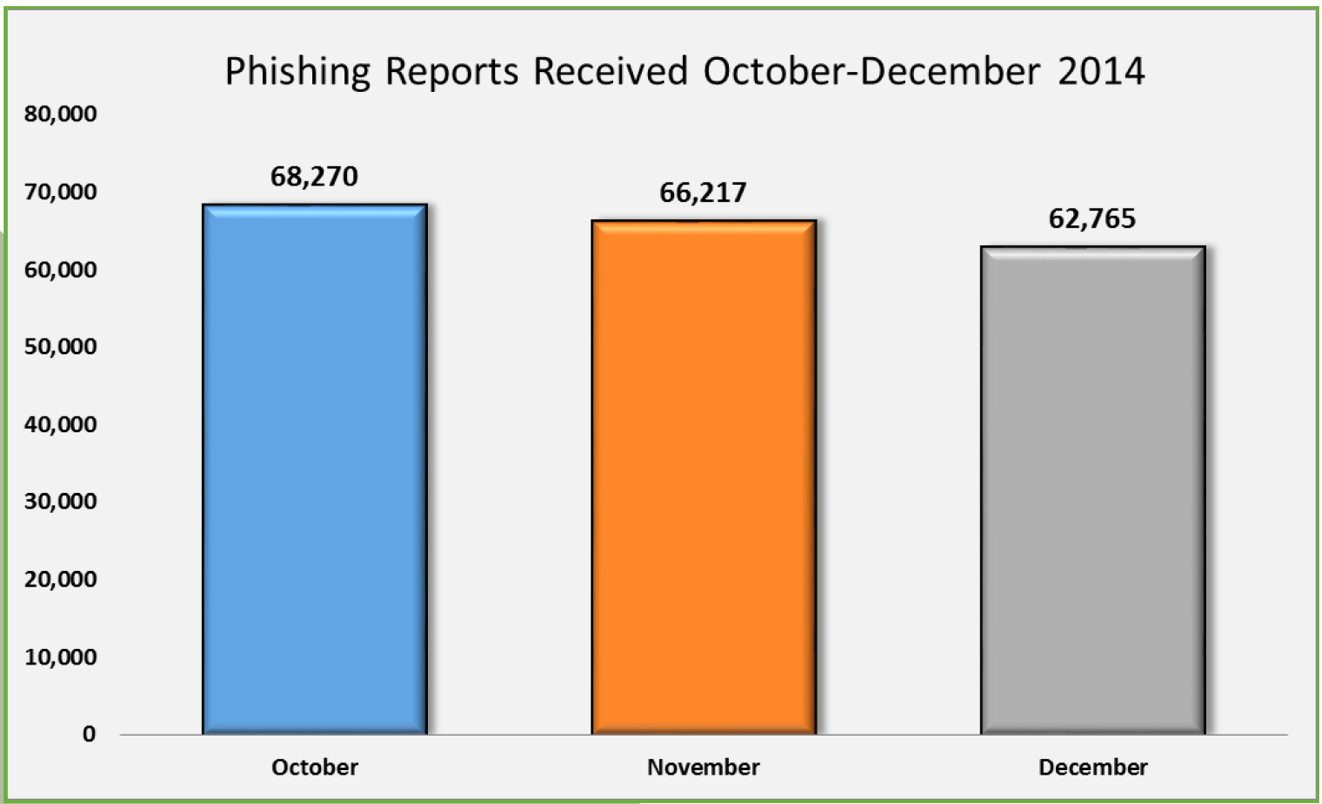
“Paypal” is on the top of list for being the victims of phishing. Paypal has been most successful way of payment for many online shopping since many years. Banks are the second most popular entities in list of attackers. Mostly all banks are targeted on the name of “Forgot Password”. Many alternatives could also be used to trap the users for their username and password of the bank account. So far many big scams have been reported. Now a day, each ecommerce websites saves your credit card and account details. Once your account credentials are revealed, it is quite easy to get the other details of your bank account. Some online shopping websites are also hot favorite like Amazon, eBay etc.



Other sectors are Emails and Social Networking. This shows that internet has been another way of taking personal revenge. Sometimes it could be teenage problems. Sometimes it could be related to your married life. In any case, the purpose is to get the personal details, to read your private emails, to reveal your secrets, to spy on you.

Number of phishing attack globally reported in the fourth quarter of 2014 is 1,97,252. This number is roughly incremented by 18 percent from third quarter of 2014. One common observation from this report is number of unique domains has been decreased from third quarter of 2014. We can say that attackers have been started using same domains for phishing attacks.

In summary, “Phishing” has been generic problem for all over the industries especially financial and ecommerce sites. This attacks have been increased from many years. Many solutions and research have been going on to solve this problems, but still it has failed to reduce the number of attacks. In fact, it has infected wide are of internet sadly. One has to find some thorough solution to solve this big problem.



***2.2 Previous and Existing Solutions***

There has been lot of research going on by different university and professional organizations to solve “Phishing” problem. Let’s take an overview of all of these approach one by one.

1. Anti-prevention (offense and defense)

This is a unique technique to handle this problem. This approach is opposite to conventional prevention. It does not restrict users to access the phishing site. It let the users access the phishing websites. This approach is applied on the top of “BlackList” approach.

When any web page is identified as a phishing web page by browser’s built in anti-phishing protection, this filter is triggered. If the user ignores the warning and continue browsing the same page, this filter intercepts the victim’s response and fill out the bogus data in the response. It creates (S – 1) bogus credentials and hide the real one between S credentials. Within a few milliseconds, those S credentials are sent to the phishing website. Same action is taken when user does not ignore the warning. But in this case, total (S) bogus credentials are being generated instead of (S - 1).

Overall aim of this solution is to confuse the phisher with the overwhelmed responses from victim’s machine. It would be difficult task for phisher to filter out the real credentials amongst thousands of data.

Limitation of this method is it assumes that phisher does not have access to query the directly to the targeted website for the validation of any individual username. This might not be the case for email based services and many other social networking services. For financial websites this could be very effective.

1. Heuristic

Heuristic based approach is the widely accepted and more popular now a day. Many research papers have been presented based on heuristic. There are mainly two components in this approach. First component fetches the require property from the webpage and second component is classifier algorithm to classify given data and make some prediction for the given input. Let’s explorer all of those heuristic one bye one.

So far most accurate heuristic is referenced search engines. The given webpage is scanned and many terms are extracted. Those terms are sent to one or more search engines like Yahoo, Google and Bing. That webpage is declared safe by heuristic if the domain name is in top N search results. However, it suffers from performance issue because of the total time taken by the round trip to search the terms.

In summary, this approach is more attractive than any other. We have worked on this approach in this paper. If you can click the right features and differentiate between legitimate websites and spoofed website then It can give you reasonably good accuracy with good performance.

1. Visual Similarity

As name itself suggest the approach of this method, it is related to comparison between visual appearance of the website. Though we think that phishing website and original website look similar in terms of design and look, sometimes there are significant difference.

One way to differentiate is based on favicon of the webpage URL. This algorithm compares the original favicon and the spoofed webpage URL’s favicon using image processing technique. This algorithm gives some value. If this value is above, then some threshold value then that URL can be detected as phishing website. Problem with this algorithm is, if the website does not have any favicon then it fails.

Second solution was proposed to decompose the webpage in to block region. Visual similarity of two webpage is then evaluated in three matric. Those are block level similarity, layout similarity and overall style similarity. If any one of these matric has got higher value then some threshold value then website is reported as phishing website

1. Blacklist

“Beyond Blacklist” is the most efficient method to detect phishing website. It collects the list of verified phishing websites from all over the world. Whenever user visits the phishing websites, it searches in to the database and responds. It gives minimum false positive rate, but it fails to detect fresh phishing websites. It has got time window problem. Since average time to be added in blacklist is not significantly long, it could be highly destructive.

1. Content Based

Most effective implementation in this category is “CANTINA”. CANTINA is based on TF-IDF information retrieval algorithm. It extracts the content of the webpage to decide weather websites is phishing or not like URL, domain name of the website.

It uses the search engine results since it uses information retrieval techniques and “Robust Hyperlinks” algorithm to improve the broken hyperlinks. Author has claimed that this algorithm achieves approximately 95% accuracy to detect phishing websites. They also suggested to use CANTINA along with heuristic to reduce false positive.

There are certain downsides of this algorithm as well. This approach is heavily dependent on searching algorithm, if phishers has achieved good page rank for the phishing websites in major search engine then it might be difficult to detect the phishing websites. Second issue is with performance as it sends query to search engine and process it based on page rank. Language is also another constraint of this phishing techniques.

We saw all the major solutions which has been proposed so far and are being used to detect phishing websites. Lot of progress has made in couple of years. But one of them has three properties like performance, accuracy and large scope of detection at the same time. Approach, suggested by us, has all of these three features.

***Chapter 3***

***Proposed Solution***

***3.1 Outline of the solution***

discuss the approach of the solution in summary

explain it with small example

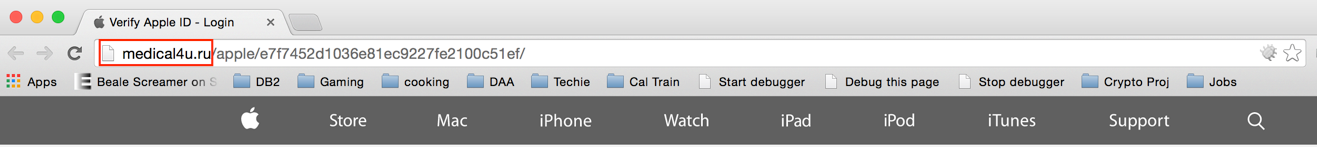
Our interactive filter adopts the heuristic based approach to detect the phishing websites. Disparity between the website’s true identity and observed identity is the main feature of this filter.

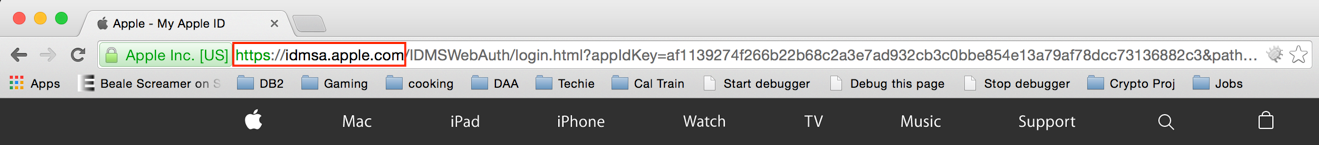
*Observed Identity:* The features which are the part of web page content and can be spoofed easily.

*True Identity:* The features which replicates the real identity of the website and very much difficult to spoofed.

Frequent terms, source domains of iFrames and source of images are the Observed identities. On the other hand, host domain is the true identity. Disparity between these two identities is measured by their textual relevance. Unique thing about this approach is to find the intersection of those characteristic which are also efficient. The reason of this uniqueness is prior heuristics never believed that there would be some common characteristics between phishing websites. It is not easy to find such disparity as it could be hidden in any form of the webpages. With the help of textual relevance algorithm, it becomes easy to find such disparities. We can set some threshold value to detect the phishing websites. Users are warned for such suspected websites. It is their decision whether they want to visit such websites or not. Through the human interaction, our heuristic can gain more accurate knowledge to improve its decision capacity. No automatic heuristic can achieve excellent detection accuracy. That is the major reason that our heuristic interacts with human to validate and update its knowledge if required. In fact, it is the best way to avoid false positive. Our results has proved our point and achieves significant good accuracy and less overhead.

Let’s take a small example to understand the disparity between two identities. We will compare it with original websites.





In the example both websites claim to be Apple web site. Let’s review both websites. In the figure 5, we can see the observed identities such as “Verify Apple ID” from its title and “apple.com” from its image/anchor domains in the web page content. If you observe, then the textual relevance between title of the website and domain of the images and anchor tag is very high. So we can say that textual relevance between observed identities is very high. Now let’s take an observed identities and true identities for comparison. True identity is its domain name in URL which is “medical4u.ru”. You can observe very less textual relevance between true identity and observed identities (medical4u.ru and Apple, medical4u.ru and apple.com). Now for figure 6, observed identities are “apple.com” and “Apple – My Apple ID” from its respective title and domain names of images. It has got higher textual relevance between true identity(apple.com) from URL. All the content of this webpage has the same origin compare to WHOIS record and copyright holder.

By seeing an example, we can say that there is higher textual relevance between observed identities for fake website whereas less textual relevance between true identities and observed identities. With this set of settings, we can easily differentiate between phishing website and benign website.

***3.2 Component of the solution***

URL features

Content features

User interaction for final decision

We divided all features in two groups majorly. One is URL features and second is Content features. Let’s see the definition and understanding of each features in detail. It would be easy to understand the classification scheme with better understanding of these features.

***3.2.1 URL features***

URL O

***3.3 Classification Scheme***

Diagram

Explain in words

Reference:

1. <https://securelist.com/analysis/quarterly-spam-reports/69932/spam-and-phishing-in-the-first-quarter-of-2015/>