Can Phishing Education Prevent Phishing Attacks?

Submitted as part of the requirements for:

CE902 Professional Practice and Research Methodology

Name: DeShaun Ormond

Tutor: Dr. Lina Barakat

Date: 26/08/19

Abstract

Phishing is currently one of the most commonly used cybercrimes that individuals face around the world. Research has been done by academia as well as companies and corporations to find a way to stop phishing, however phishing continues to increase daily. This study sets out to determine if taking an educational approach is an effective way to combat phishing, can phishing education prevent phishing attacks? A google chrome extension was created to educate the user on what to look for in URLs for signs of phishing and actively scan each website browsed for phishing. Analysis of the extension determined educating individuals on the dangers of phishing and what to look for, can help with prevention of phishing. These results indicate that education should play a key role in phishing prevention. It is recommended that organizations implement more educational tools to combat phishing. Further research is required to identify other factors that can strengthen the effectiveness of education.

Keywords: Social engineering, Phishing, Cyber Security, Machine Learning, Random Forest

Table of Contents

Introduction

Criminals online or cyber criminals partake in many cyber-attacks, however undoubtably the most popular cyber-attack is phishing. Phishing is the art of deceiving someone into giving them their personal information, "typically in the form of usernames, passwords, credit card numbers, bank account information or other important data in order to utilize or sell the stolen information"[1]. The cost of a phishing attack can vary depending on the goal of the attack. A phishing attack can cost a mid-size company 1.6 million dollars, in 2017 Google and Facebook lost 77 million pounds to a single phishing attack[2][3]. With these attacks a lot of people have sought out to find an answer to the question, how can we stop phishing attacks? This question has been examined heavily by researchers worldwide. Since the rise of phishing attacks, there has been multiple attempts to find ways that will solve this question, from actively searching and blocking bad websites behind the scenes to in-person educational courses on what to look for. However, with more phishing attacks being sought out everyday and becoming more sophisticated in deception, these methods alone cannot keep up. Data from [4] reports a 76 percent increase among phishing attacks in 2019 compared to attacks in 2017.

This paper answers the question with a software designed to educate and defend. We created a google chrome extension that is capable of detecting phishing attacks at a rate of 75% and is able to effectively educate users on what to look for. The software incorporates a random forest model trained on data from [0] to detect 17 key features in a websites URL that are commonly present in phishing URLs. The extension aims to educate users with its many components, it includes randomly generated facts about phishing, an external educational page, and allows the chance for users to submit inquiries about potential new undetected phishing sites.

We gathered 10 participants and conducted a user test study to test how users would react to the app and its components. We also asked 22 participants to fill out a survey on internet safety in order to gather common knowledge on the subject of phishing. The key findings are:

- The survey showed:
    - Only 9% participants check to see if the website matches the URL
    - 36% participants have a considerate knowledge of what phishing is
    - 50% practice internet safety and 63% hover over links before clicking on them.
- The user study showed:
    - 80% of participants heeded the blocked page warning and closed the dangerous site.
    - All participants presented with the blocked page warning were able to understand and make a decision based off of the information provided
    - Phishing language was hard to comprehend and would lead to a loss of interest.

The results of these experiments gave insight into key improvements that needed to be made before we allowed it to market. We implemented a quiz for the users to stay engaged an continue learning, a demonstration icon so users can know what to expect and their options when the software detects a phishing site and a slideshow that is able to clearly define and present the user with the results of the 17 features that the software looks for.

Related Work

The first phishing attack though not considered at the time, can be traced back to 1995 [5]. For over 24 years phishing has grown rapidly and become a greater security threat worldwide. With this growth many individuals have created tools, wrote books, researched and entered career fields in order to produce a solution to this problem. During our research we came across a number of anti-phishing tools and education research that closely relate to our goal of this project.

[6], [7], and [8] are phishing detection mechanisms that were developed all with the similar goal of stopping phishing attacks. [6] analyzed the phishing website detection across a number of machine learning models including logistic regression, random forest, and support vector machine. [7] built an extension that performs web-crawling, a technique that checks the contents of the page for certain phishing characteristics. [8] is a lite chrome plugin that shows the user, what features a website has in the form of visual representation. These three extensions all yield a random forest model classifier. With this along with independent research we chose this to be the underlying machine learning model the software will have in order to detect newly browsed URLs.

In the education field there has been a number of reports, documents, and studies carried out on phishing. [9] is a study that investigated if anti-phishing tools in the browser at the time were effective in stopping phishing techniques. [9] gathered a list of notable anti-phishing tools and tested each one to examine their efficiency. They gathered that the source of data that was trained into the algorithm significantly impacts the results of the anti-phishing tool. This finding was important to note for this project, if our extension was going to educate the users it would have to demonstrate a high detectability efficiency rate on URLs both phishing and legitimate.

[10] examined the philosophical question of 'why phishing works?'. In their study they exposed 22 participants to 20 websites and asked them to determine the authenticity of the site. Their overall aim was to gain data to provide evidence on which strategies were the most successful at deceiving general uses. Their results were that '23% of participants did not look at browser-based cues such as the address bar, status bar and the security indicators, leading to incorrect choices 40% of the time ' [10]. From this study we drew a conclusion that redirecting the user to an information page detailing information about the site they requested and the dangers of it is the optimal way of overcoming these problems.

Our research also displayed that educating users on phishing has been explored. [11] introduces the idea of having a 'change of direction' in security awareness. He proposes that developers should think like general users and how they would make decisions in their everyday lives rather than flooding users with information. A quick search on the chrome webstore for 'phishing' will produce a number of phishing extensions available for download. However, none of them applied the idea to educate the user on what to look for. A study done by [12] aimed to evaluate a number of website security tools, they conducted an experiment and it was reported that '53 percent of their study participants still attempted to log in to a site after their task was interrupted by a strong security warning' [11]. This information proves critical that by only alerting the user on potential phishing sites is not adequate enough. Education is required in order to reduce the number of attacks and breaches worldwide.

Background

Those who seek to exploit human vulnerabilities or security vulnerabilities online for their own personal malicious gain are considered cyber criminals [13]. With the increase in online activity and technology advancement, comes along the increase in cyber criminals and cyber-attacks on individuals and companies globally. One of the most common of attacks carried out by those criminals is Phishing.

"Phishing is a type of Internet fraud that seeks to acquire a user's credentials by deception" [14]. Those who perform this attack are called phishers and they will generally contact an individual posing as someone trustworthy to gain confidential information like passwords, credit card numbers or bank account details. Contact from phishers will be in either the form of emails or ads that will lead you to a spoofing website, a fake website that looks similar to a popular website. in the hopes of deceiving you into entering your personal information. A visual representation of a phishing attack can be found in figure 1. Phishing is the preferred type of attack for cyber criminals with 91% of all cyber-attacks beginning with phishing [15]. Social engineering plays a key part in these attacks.

Social engineering is the act of manipulating a person to take an action that may or not be in their best interest [16]. [17] explains that humans are predisposed to trust people and if they have not been given a reason to not trust, there will be no need to second guess giving your information to a website that looks official. Essentially cyber criminals use social engineering to exploits human vulnerabilities to further their malicious agenda and since all humans make mistakes, they have plenty of chances to succeed.

The number of phishing attacks grew 40.9% in 2018 and an estimated 135 million phishing attacks are attempted every day [18][19]. This means to stop this machine learning will have to be incorporated with any security measure to have a fighting chance of stopping attacks.

Machine learning is a method of data analysis that allows for systems to learn from data, identify patterns and make decisions with minimal human intervention [20]. This project incorporates a trained machine learning model to classify URLs as phishing or legitimate when a user navigates to a webpage.

Research Question

The overall aim of this paper is to fill a key gap in common knowledge about phishing. This paper examines the idea that if individuals were educated on what to look for in a websites URL then those individuals would be better equipped to detect phishing websites if encountered. Achieving this would lead to less security breaches as well as less money being spent by an individual to defend or recuperate from phishing attacks.

To evaluate this question a phishing detection system was created and aimed to determine the legitimacy of a website when browsed by the user. This software incorporates a machine learning algorithm that uses 17 features and previously trained data from [0] to be able to accurately predict future webpages when visited. If the system detects a phishing URL then the extension will promptly alert the user by redirecting them to an information page. This page will instruct the user on why they are seeing this and show them options on how to proceed. The overall objective of this software is to be able to prevent phishing websites while educating users on what to look for in a webpage URL.

Methodology

In an effort to educate individuals on what to look for in a URL, an implementation of a google chrome extension is the desired choice. “With over 4 billion people actively online and with more individuals using google chrome than any other browser”, google chrome allows for our extension to reach and educate a wide variety of individuals [21]. Google chrome allows developers to create a small software program that enhances a user’s browsing experience, an extension [22]. This provides a wide variety of opportunities to educate the user on phishing prevention as well as seamless URL scanning to detect phishing websites.

To develop this software a number of resources was needed. The software needed to incorporate a machine learning model, data of known phishing and legitimate URLs and an educational plan on how to convey information to users. Experiments were also needed in order to determine the efficiency and success of the system, software and user testing.

Design and Analysis

Datasets

The datasets and features derived from [0] contains 1,105 URLs and 30 features. [0] compiled 4,989 phishing and 6,157 legitimate URLs mainly from PhishTank’s archive, MillerSmiles’ archive and Googleâ€™s, major websites that contain data on phishing scams [0]. This dataset includes 30 features that are applied to each website to determine if that feature is present in the current URL. [0] states that each feature included has proven to be sound and effective in predicting phishing websites. We decided to analyze each feature and determine its level of importance in affecting the outcome of the predictability, shown in figure 2. A further look into the features revealed that there was a need to drop some features from the dataset due to the inability to execute the feature on a website within a reasonable timeframe and the level of importance.

After the removing features, we were left with 17 applicable features, shown in figure 3. While each feature looks for a phishing characteristic, [0] categorizes them into three main categories, address bar-based features, abnormal based features, HTML and JavaScript based features.

The address bar-based features are focused on the URL address of the webpage. They obtain the current URL, examines the content, applies a rule-based logic of phishing or no phishing to determine its legitimacy. The HTML and JavaScript based features focus on internal content of the page and check for known phishing tendencies that can be found in HTML and JavaScript code. The abnormal based features scan the webpages internal content and looks for an abnormal amount of code, excessive use of HTML or JavaScript code can likely mean the webpage is trying to deceive or confuse an individual.

The top three features that have the most importance in determining predictability are analyzing https, sub domains, and anchor tag checking.

- SSL Final State(https)

This feature requires checking if the website uses https and if the age of the certificate. Https is short for Hyper Text Transfer Protocol and is an indicator that the website is secure and communicating privately. In order to achieve this, developers need to gain a certificate from certified internet authorities. Unfortunately obtaining a certificate is not complex and any cybercriminal can acquire one. That is why this feature also

analyzes the age of the certificate. Research concluded by [0] shows that the minimum age for a https certificate relating to an authentic website is two years. Therefore, any website that does not meet this requirement will be considered phishing.

- Having_Sub_Domain

When an individual register a URL they have to pick a name for their website. This is known as a domain name and is a unique identifier to each website. A domain name like example.com. A domain name is the name of a website, example.com. Domain names are unique identifiers that lead a user to a registered site. A domain name is made up of two components, a country code(.uk, .com) and a second level domain(example, google). While cyber criminals cannot use names for websites that have already been register by a company, they can add subdomains to it to fool users. An example of adding subdomains to example.com would be login.accounts.uk.example.com. The number of dots in a URL indicate the number of subdomains. If a URL has more than 2 dots in the domain then his feature will indicate phishing.

- URL of Anchor

[0] incorporated a feature that checks the contents of the page, web-crawling. This feature crawls through the content of the webpage to look for an anchor tag. An anchor tag is an HTML element that allows developers to redirect users to an external URL. A normal page will contain around 31% to 67 % of anchor tags in its webpage. However, webpages with more than 67% of anchor tags likely indicate that the webpages main goal is to redirect you another webpage instead of exploring its own page.

All the features in each category are applied to each website browsed and the results are added up and then multiplied by the weights of the algorithm to determine the legitimacy of the site. With the majority of features haling from the address bar-based category,_ it was decided to give preference to these features and center the extension around educating users on how to detect and determine legitimacy of it.

Algorithm

The goal of the software is to be able to correctly identify between phishing and non-phishing URLs. Since the outcome is clearly defined as legitimate or phishing, it was evident that this is a classification problem. The software additionally required knowledge on phishing URLs to accurately detect incoming phishing URLs. With these requirements outlined, we researched and established that a machine learning supervised model approach would be optimal for this problem. The machine learning model would allow us to train the computer on data from [0] and then program it and then allow the software to classify incoming URLs without human intervention. To

We gathered seven supervised model approaches and translated the algorithm to python code in order to compare on our dataset. Python coding language was selected because of the extensive set of libraries it has for machine learning. The scientific library chosen to create multiple machine learning models was scikit-learn, a tool for data mining and data analysis [28]. Each machine learning model has parameters that are able to improve the accuracy of the model. Scikit-learn also allowed us to Grid SearchCv to determine the optimal parameters[30]. Once we obtained the parameters, we trained the model on 70% of the data, predicted on the remaining 30%, and recorded outcomes.

Shown in figure 4, eight models were able to perform well and achieve an accuracy higher than 90%. A closer look at the models and four models achieved an accuracy of 94%, random forest, support vector machine, decision tree and neural network. We eliminated the other four from comparison and performed tests with the remaining models. The outcome of the test as shown in figure 5, signify that using a random forest model will grant us a slight advantage in accuracy over other models.

A random forest model in its simplest form is a larger and more powerful decision tree model. A decision model is a "series of yes/no questions asked about the dataset eventually leading to a predicted class"[26]. Figure 6 illustrates how a decision tree model would encounter a node that contains a question about the data, this continues until a predictive value has been reached. Our algorithm uses the random forest model which is made up of multiple decision trees, figure 7. Instead of having one decision tree predicting an outcome, the random forest model combines hundreds, or thousands of decision trees and the final prediction is made by averaging the predictions of each individual tree[26].

Once the model was selected and tested, we ranked each feature on how much influence they had on predicting the outcome(weight of importance). We then extracted the percentage in forms of decimals and documented it for later use.

Educational Plan

To effectively communicate complex information about phishing to potential uninformed users, a list of necessities was created. This list was our baseline in the approach we took to present information through the app. The outline included:

- Keep It Short
- Include Facts
- Appeal to all types of learning.
- Interactive material

A popular mantra for giving information is KISS, Keep It Short and Simple. This mantra will allow us to effectively deliver quick and precise information to the user without any confusion to the user. The information presented will need to be accurate and not applied knowledge. For this reason, we use information provided by two reputable sources, [5] and [27]. It is also important for the extension to keep the user engaged in the app when a user clicks on the app or its components. The extension should also appeal to the three types of learning: visual, auditory and Kinesthetic.

Implementation

After the algorithm was completed, we began to implement it to the extension. Due to coding rules set by Google, coding for the extension could only be done in HTML and JavaScript. Therefore, we coded the 17 features found in [0] in JavaScript. The outcome of the features could only be two options, 1(phishing) or -1(not phishing), and the results would then be multiplied by the weight of importance. This would output a number that determines the legitimacy of the website.

The educational plan was implemented throughout the extension's components. The extension has three components installed that incorporate the educational plan.

- Random generated Facts

  When a user clicks on the extension one of twenty facts will be displayed on the front page. These facts include details about URLs, rise of phishing and internet safety tips.
- Learn More Link

  Located at the bottom of the extension is a “Want to Learn More” link. When clicked, this link will navigate users to a custom-built page for users to explore, shown in figure 8 and 8.1. The page contains a six of tabs that provide useful information on phishing 4 links to external games where users can play to test their phishing knowledge. These tabs include:
  - Structure tab

    Detailed information about the structure of a URL and components that make up a URL.
  - Summary Tab

    A simple walkthrough of phishing that answers the common questions asked about phishing. displayed are six questions and six bullet points that clearly state the definition of phishing, who is behind it, what to do, when to expect it, where to expect it, why it is happening and how to defend from it.
  - Example Tab

    This tab displays two Facebook login pages and asks the user to determine the non-phishing sign in page. When the page is clicked the correct answer will appear on the page along with a message relating to the deception that phishers use.
  - Lookout Tab

    An informative tab that instructs the user on what to look for in order to recognize and prevent phishing.
  - Statistics Tab

    Eight facts pertaining to phishing and the effect is has had on the world is displayed in bullet point form for the users.
  - Videos

    Four links to external videos that give a visual overview of phishing. This tab allows visual learners to easily processes information given.
- Fire Symbol Icon

  This icon is displayed on the top right corner of the extension. When clicked the user will be navigated to a custom page. This page is what the user will see when they the extension detects a phishing website. Shown in figure 9 the page informs the user on what this page’s intentions are and why it is being shown. The page shows the users requested URL on the screen and presents three options for the user. The user can navigate back to their previous website, navigate to the learn more page or continue to the desired page. This icon allows for the user to know what to expect and how to react to the page.
- Quiz Icon

We implemented a quiz for the user to test their skills and keep them engaged with the app. Figure 10 shows what the user will see when the icon is clicked. The quiz presents five questions to the user relating to phishing information. The user will have 10 seconds to select the correct answer between three choices given, if the user does not select an answer before the time is up then that answer will be marked as incorrect. At the end of the quiz the user will be shown their score with a personal message relating to their score.

- List Icon

This icon will display the outcome of every feature to the user. We simplified the information for better understanding and displayed a green check or red-x next to a message describing the results. The green check will be present when the phishing characteristic has not been found, and the red-x symbol will be displayed when it has been found.

- Report Button

A button is presented to the user at the bottom of the extension that allows for the user to report a phishing page. When the button is clicked the extension will display a message informing the user that this URL has been documented, shown in figure 11. The extension will display the current tab to the user and send the URL to a list for us to verify.

Evaluation

The following sections will outline steps taken to conduct each experiment and the results of those experiments. A successful evaluation of software experiments will be an accuracy of 80% or more detectability with phishing websites and legitimate sites. The user experiments will determine the extension's simplicity, usability and the visual learning implemented.

Software Testing

The data and features provided by [0] when trained and tested using a random forest model, produces a 94% accuracy classification rate. To determine the effectiveness of classifying on real-time phishing websites an experiment was carried out. The experiment consisted of presenting the software with 100 websites both phishing and legitimate provided by [21], [22] and [23].

The results of this experiment resulted in a 98% accuracy rate of correctly identifying legitimate sites and a 75% accuracy rate on detecting phishing site.

User Experiments

10 participants were gathered to test the functionality and responsiveness of the app. We outlined an experiment to test the functionality and responsiveness of the app. We gathered 10 participants and provided them a list of instructions. Each participant was instructed to visit five websites that would be provided, log on with credentials provided by us and complete an action on the site. As shown in figure #, the sites were popular social media sites and they were instructed to post or follow an individual of their choosing. Out of the given five websites to visit, unbeknown to the participant three of the sites were designated phishing sites. The phishing sites were made using a phishing toolkit from [29]. Figure 12 shows the resemblance of our Facebook phishing page compared to the real phishing page and the potential for misleading a participant.

Throughout the experiment the extension would be install and perform its analyzing of webpages in the background. When the participant navigated to two of the three phishing sites the extension would detect it, alert the user with a custom-made site that would instruct the user about potential phishing site and on how to proceed. For the third phishing site only, a visual indicator was given, the extension icon would change from a smiley face to a red mad face. Throughout the test we recorded each user experience and their actions and refrained from giving participants help when they were faced with a blocked page warning.

The results of the experiment are as follows:

- All participants acknowledged that they encountered the blocked page site twice throughout the experiment and were able to clearly understand their options.
- 3/10 participants bypassed the blocked page warning and continued to the phishing website
- 6/10 participants conveyed that the language on the extension was difficult to grasp and needed to be broken down more.
- 3/10 participants admitted to noticing the visual indicator change on the third phishing site and exited the website.

Survey

After we concluded the user test, we began working on updating the extension and its inefficiencies. We decided that we needed to create a survey to obtain a collective measure of how individuals are acting online. This would allow us to acquire real-time information on the common internet safety practices individuals are following.

22 participants were asked 9 questions that would give insight on whether they performed common internet safety techniques. These results help get a grasp on what individuals are already aware of and practice, as well as help us model the app in the future to include common internet safety techniques that individuals less likely do. Each question presented to the user allowed us to consider what topic to focus on the most. Figure 14 illustrates each survey question in a graphical pie chart along with the percentage of answers provided.

Our survey revealed that while participants have heard of phishing before and do often perform common internet safety techniques, there is a significant void in the practice of internet safety. The survey questions reveal compelling findings on what individuals are aware of.

- Contrary to data from [21] only 40% of the participants admitted to using google chrome, with 50% stating they use safari and 9% using Firefox.

When presented with questions about their online activity presence,

- 50% of the participants admitted to always being online.
- 64% said that they tend to overshare information
- 55% acknowledge that they tendency to share information with others online.

When participants were asked about performing internet safety techniques,

- 64% said they hover over links before visiting them,
- 9% of admitted to always looking to see if the website matches the URL.

Project Plan

Outline of this project's goals, objectives, tasks, and milestones that were planned out for the purpose of completing this project is included in the Appendix Section, figure 14 .

Limitations

There are a number of ways that a cybercriminal can contact an individual and perform a phishing attack. Most happen through emails that lead them to a phishing website or an email that will download dangerous files when clicked on. This project focuses only on phishing websites, by scanning the URL for features and data of common phishing URLs provided by [0].

It should be noted that the participants selected for the study and the survey can be categorized into the age group of 18-23 and the results from this survey are unable to include wide variety of ages.

While the extension does detect phishing websites at a rate of 75%, it is not meant to stand alone or outperform other security systems produced by companies or corporations. To achieve a high level of detection accuracy using machine learning, the extensions would require continuous updating of data, adding new phishing features and an increase in the number of data entry points. This extension is best categorized as an educational tool, a tool that informs the user on common phishing tendencies to look for in a website's URL.

Conclusion

This paper investigates if education will be able to prevent phishing attacks. We created a google chrome extension with the aims of educating users on how to detect phishing URLs by using well-known phishing URL tendencies . The extension contained a random forest algorithm, examined new webpages for 17 phishing features and contained five components to effectively convey and educate the user. We conducted a software accuracy test for the algorithm and achieved a 75% accuracy rating among newly detected phishing websites. A user study was carried out to determine the efficiency of the extension and the usability. Findings from the study revealed that quick and precise information along with components that are able to keep the user engaged are required to effectively educate users on phishing attacks. We conclude that an educational approach can prove beneficial to preventing phishing attacks and recommend that education be the preferred tool organizations use to combat phishing.

Further research is required to investigate top factors that need to be included in an educational approach. information provided from user testing and surveys show that research will need to be carried out on how to properly present phishing information to the users along with more educational components to keep the user engaged. More work is also to be considered in order to achieve a desired detection rate of 90% or higher.

References

[0]D. Due and C. Graff, "Phishing Websites Data Set", *Archive.ics.uci.edu*, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/phishing+websites. [Accessed: 20- Jul- 2019].

[1]"What Is A Phishing Attack?", *Cloudflare*, 2019. [Online]. Available: https://www.cloudflare.com/learning/security/threats/phishing-attack/. [Accessed: 09- Aug- 2019].

[2]E. Katz, "Phishing Statistics: What Every Business Needs to Know", *dashlane blog*, 2019. [Online]. Available: https://blog.dashlane.com/phishing-statistics/. [Accessed: 09- Aug- 2019].

[3]M. Samarati, "Phishing scam cost Google and Facebook £77m - IT Governance Blog", *IT Governance Blog*, 2019. [Online]. Available: https://www.itgovernance.co.uk/blog/phishing-scam-cost-google-and-facebook-77m. [Accessed: 09- Aug- 2019].

[4]A. Moscaritolo, "Beware: Phishing Attacks Are on the Rise", *PCMAG*, 2019. [Online]. Available: https://www.pcmag.com/news/366210/beware-phishing-attacks-are-on-the-rise. [Accessed: 09- Aug- 2019].

[5]"Phishing | History of Phishing", *Phishing.org*, 2019. [Online]. Available: http://www.phishing.org/history-of-phishing. [Accessed: 24- Jul- 2019].

[6]A. Dobhal, "abhishekdid/detecting-phishing-websites", *GitHub*, 2019. [Online]. Available: https://github.com/abhishekdid/detecting-phishing-websites. [Accessed: 24- Jul- 2019].

[7]R. Naik, "philomathic-guy/Malicious-Web-Content-Detection-Using-Machine-Learning", *GitHub*, 2019. [Online]. Available: https://github.com/philomathic-guy/Malicious-Web-Content-Detection-Using-Machine-Learning. [Accessed: 24- Jul- 2019].

[8]"picopalette/phishing-detection-plugin", *GitHub*, 2019. [Online]. Available: https://github.com/picopalette/phishing-detection-plugin. [Accessed: 24- Jul- 2019].RW5 = [?]Y. Zhang, S. Egelman, L. Cranorand J. Hong, “Phinding Phish: Evaluating Anti-Phishing Tools”. figshare, 29-Jun-2018 [Online]. Available: https://kilthub.cmu.edu/articles/Phinding_Phish_Evaluating_Anti-Phishing_Tools/6470321/1. [Accessed: 24-Jul-2019]

[9]Y. Zhang, S. Egelman, L. Cranorand J. Hong, “Phinding Phish: Evaluating Anti-Phishing Tools”. figshare, 29-Jun-2018 [Online]. Available: https://kilthub.cmu.edu/articles/Phinding_Phish_Evaluating_Anti-Phishing_Tools/6470321/1. [Accessed: 10-Aug-2019]

[10]R. Dhamija, J. Tygar and M. Hearst, "Why Phishing Works", UC Berkely, 2019.

[11]I. Kirlappos and M. Sasse, "Security Education against Phishing: A modest Proposal for a Major Rethink", IEEE, London, 2012.

[12]S. Schechter, R. Dhamija, A. Ozment and I. Fischer, "The Emperor's New Security Indicators", Massachusetts Institute Laboratory, Harvard University, 2019.

[13]"Cyber crime - National Crime Agency", *Nationalcrimeagency.gov.uk*, 2019. [Online]. Available: https://nationalcrimeagency.gov.uk/what-we-do/crime-threats/cyber-crime. [Accessed: 12- Jul- 2019].

[14]"Cyber Attack - What Are Common Cyberthreats?", *Cisco*, 2019. [Online]. Available: https://www.cisco.com/c/en/us/products/security/common-cyberattacks.html. [Accessed: 12- Jul- 2019].

[15]A. Savvas, "91% of cyberattacks begin with spear phishing email", *Techworld*, 2019. [Online]. Available: https://www.techworld.com/news/security/91-of-cyberattacks-begin-with-spear-phishing-email-3413574/? [Accessed: 23- Jul- 2019].

[16] C. Hadnagy, Social engineering: The Art of Human Hacking, 1st ed. Indianapolis: Wiley Publishing, 2011.

[17]R. Kramer, "Rethinking Trust", *Harvard Business Review*, 2009. [Online]. Available: https://hbr.org/2009/06/rethinking-trust. [Accessed: 12- Jul- 2019].

2019].

[18]"2019 PHISHING TRENDS AND INTELLIGENCE REPORT The Growing Social Engineering Threat", 2019.

[19]"Ultimate Guide to Phishing", *Meta Compliance*, 2019. [Online]. Available: https://www.metacompliance.com/resources/ultimate-guide-to-phishing/. [Accessed: 12- Jul- 2019].

[20]"Machine Learning: What it is and why it matters", *Sas.com*, 2019. [Online]. Available: https://www.sas.com/en_gb/insights/analytics/machine-learning.html. [Accessed: 12- Jul- 2019].

[21]D. Ormond, "Can Education prevent Social Engineering", Masters, University of Essex, 2019.

[22] "What are extensions? - Google Chrome", Developer.chrome.com, 2019. [Online]. Available: https://developer.chrome.com/extensions. [Accessed: 21- Mar- 2019].

[23]"Alexa - Top sites", *Alexa.com*, 2019. [Online]. Available: https://www.alexa.com/topsites. [Accessed: 23- Jul- 2019].

[24]"Top Websites: The 500 Most Popular Sites on the Internet", *Moz*, 2019. [Online]. Available: https://moz.com/top500. [Accessed: 23- Jul- 2019].

[25]"OpenPhish - Phishing Intelligence", *Openphish.com*, 2019. [Online]. Available: https://openphish.com/. [Accessed: 23- Jul- 2019].

[26]W. Koehrsen, "An Implementation and Explanation of the Random Forest in Python", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76. [Accessed: 20- Aug- 2019].

[27]"How to Recognize and Avoid Phishing Scams", *Consumer Information*, 2019. [Online]. Available: https://www.consumer.ftc.gov/articles/how-recognize-and-avoid-phishing-scams. [Accessed: 20- Aug- 2019].

[28]"scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation", *Scikit-learn.org*, 2019. [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 20- Aug- 2019].

[29]"CodeExplainedRepo/Multiple-Choice-Quiz-JavaScript", *GitHub*, 2019. [Online]. Available: https://github.com/CodeExplainedRepo/Multiple-Choice-Quiz-JavaScript. [Accessed: 26- Aug- 2019].

[29]"How To Create a Slideshow", *W3schools.com*, 2019. [Online]. Available: https://www.w3schools.com/howto/howto_js_slideshow.asp. [Accessed: 26- Aug- 2019].

[30]"3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.21.3 documentation", *Scikit-learn.org*, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html. [Accessed: 26- Aug- 2019].

[31]E. Büber, "Phishing URL Detection with ML", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5. [Accessed: 26- Aug- 2019].

Appendix

Figure #1

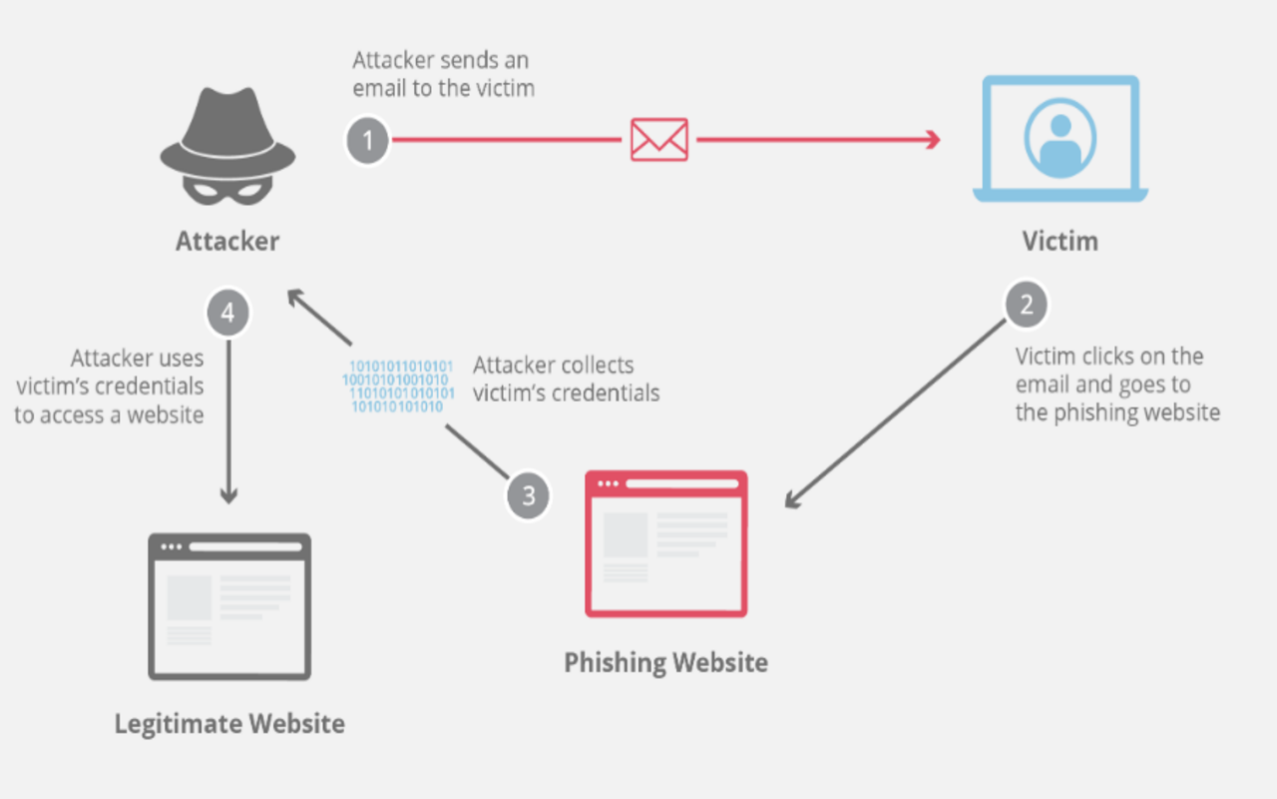


Figure #2

SSLfinal_State	0.316735
URL_of_Anchor	0.242643
web_traffic	0.074410
having_Sub_Domain	0.068941
Prefix_Suffix	0.043691
Links_in_tags	0.041233
Request_URL	0.019638
SFH	0.019430
Links_pointing_to_page	0.019266
age_of_domain	0.017108
Domain_registeration_length	0.015886
having_IP_Address	0.014836
Google_Index	0.013796
Page_Rank	0.012595
DNSRecord	0.012118
URL_Length	0.008494
Redirect	0.005899
HTTPS_token	0.005855
having_At_Symbol	0.005400
Submitting_to_email	0.005349
Shortining_Service	0.005107
popUpWidnow	0.005021
Statistical_report	0.004738
Abnormal_URL	0.004386
Favicon	0.003952
double_slash_redirecting	0.003742
on_mouseover	0.003503
port	0.002496
Iframe	0.002407
RightClick	0.001326

Figure #3

```
                            importance
SSLfinal_State                0.395925
URL_of_Anchor                 0.311015
having_Sub_Domain             0.065243
Links_in_tags                 0.051577
Prefix_Suffix                 0.049091
Request_URL                   0.025546
SFH                           0.022764
having_IP_Address             0.014109
URL_Length                    0.011979
having_At_Symbol              0.008626
HTTPS_token                   0.008372
Favicon                       0.007613
Shortining_Service            0.007138
Submitting_to_email           0.006483
double_slash_redirecting      0.006141
Iframe                        0.004785
port                          0.003594
Accuracy of Random Forest:  95.41754597527886
Running time:  0.5472254753112793
```

Figure #4

1. Random Forest = 96.85%
2. Logistic Regression = 92.29%
3. SVM-Linear = 92.40%
4. SVM = 96.74%
5. Decision Tree = 96.38%
6. KNN = 94.86%
7. Naive Bayes = 61.18%
8. DNN = 96.09 %

Figure #5

```
running random forests...
Accuracy:  94.42267108833283
Runtime:  2.36765718460083

running svmNoLinear...
Accuracy:  94.24178474525173
Runtime:  1.0642051696777344

running Decision Tree...
Accuracy:  94.15134157371119
Runtime:  0.00902700424194336

running Neural Network...
Accuracy:  93.94030750678324
Runtime:  0.9836359024047852
Overall Runtime:  4.4763102531433105

[Done] exited with code=0 in 5.908 seconds
```

Figure #6

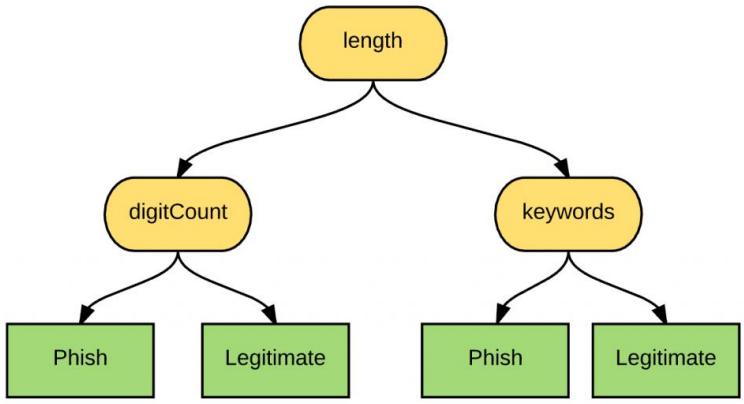


Figure #7

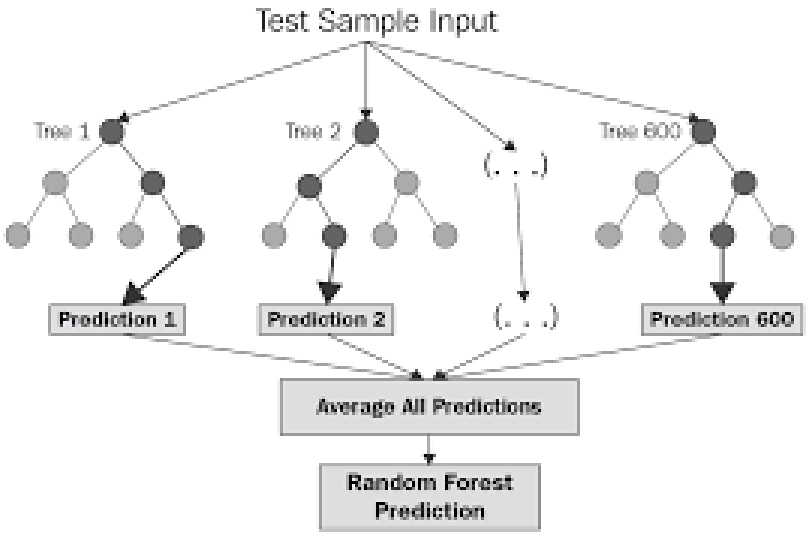


Figure #8

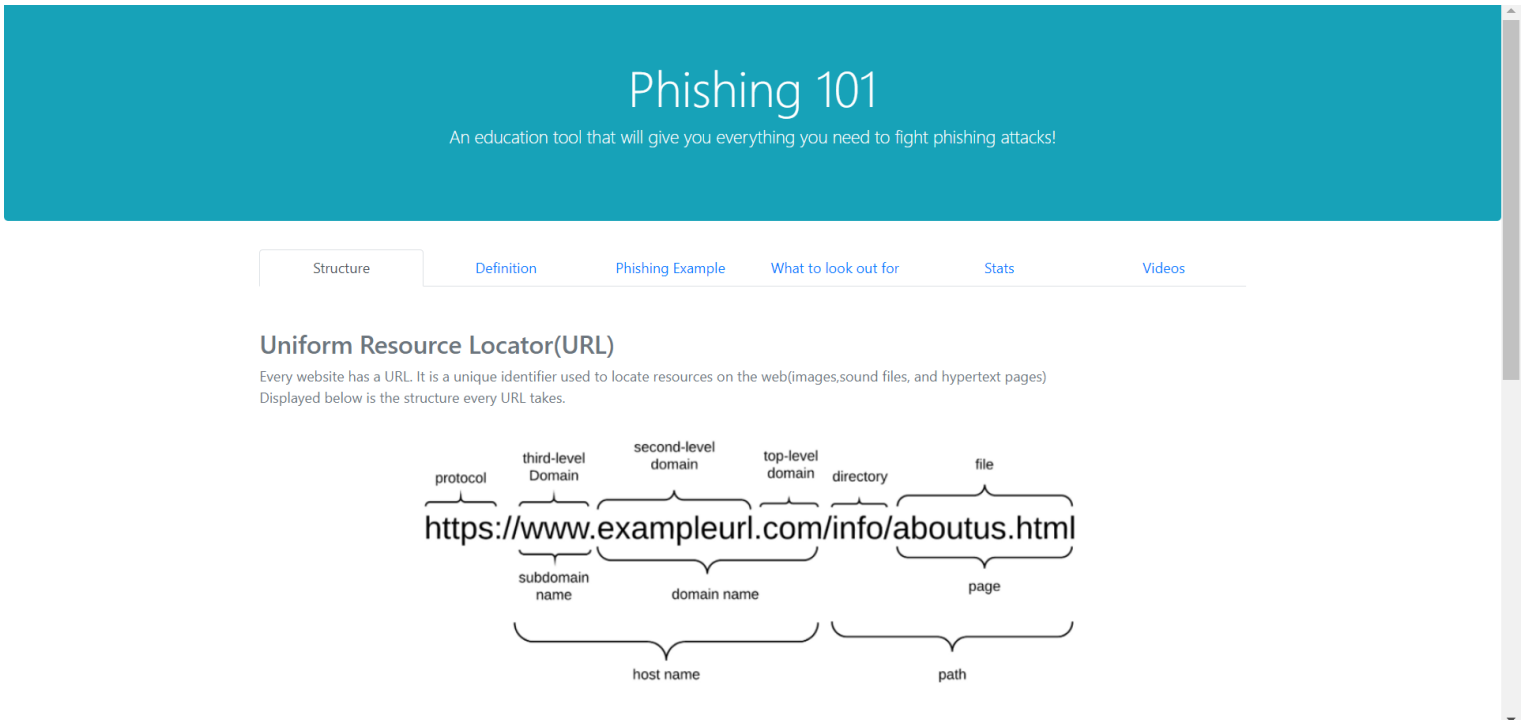

Phishing 101
An education tool that will give you everything you need to fight phishing attacks!
Structure
Definition
Phishing Example
What to look out for
Stats
Videos
Uniform Resource Locator(URL)
Every website has a URL. It is a unique identifier used to locate resources on the web(images,sound files, and hypertext pages)
Displayed below is the structure every URL takes.
protocol
third-level Domain
second-level domain
top-level domain
directory
file
https://www.exampleurl.com/info/aboutus.html
subdomain name
domain name
page
host name
path

Figure #8.1

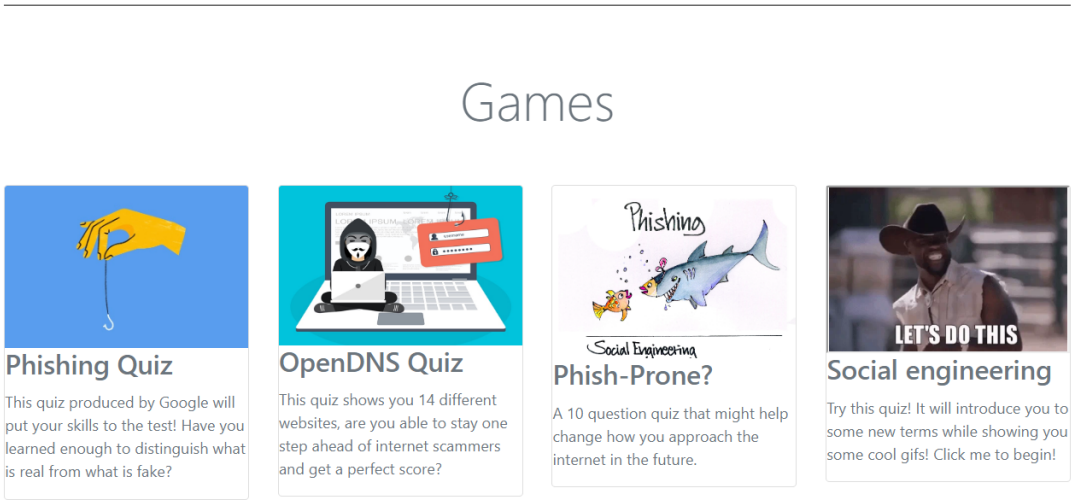


Figure #9

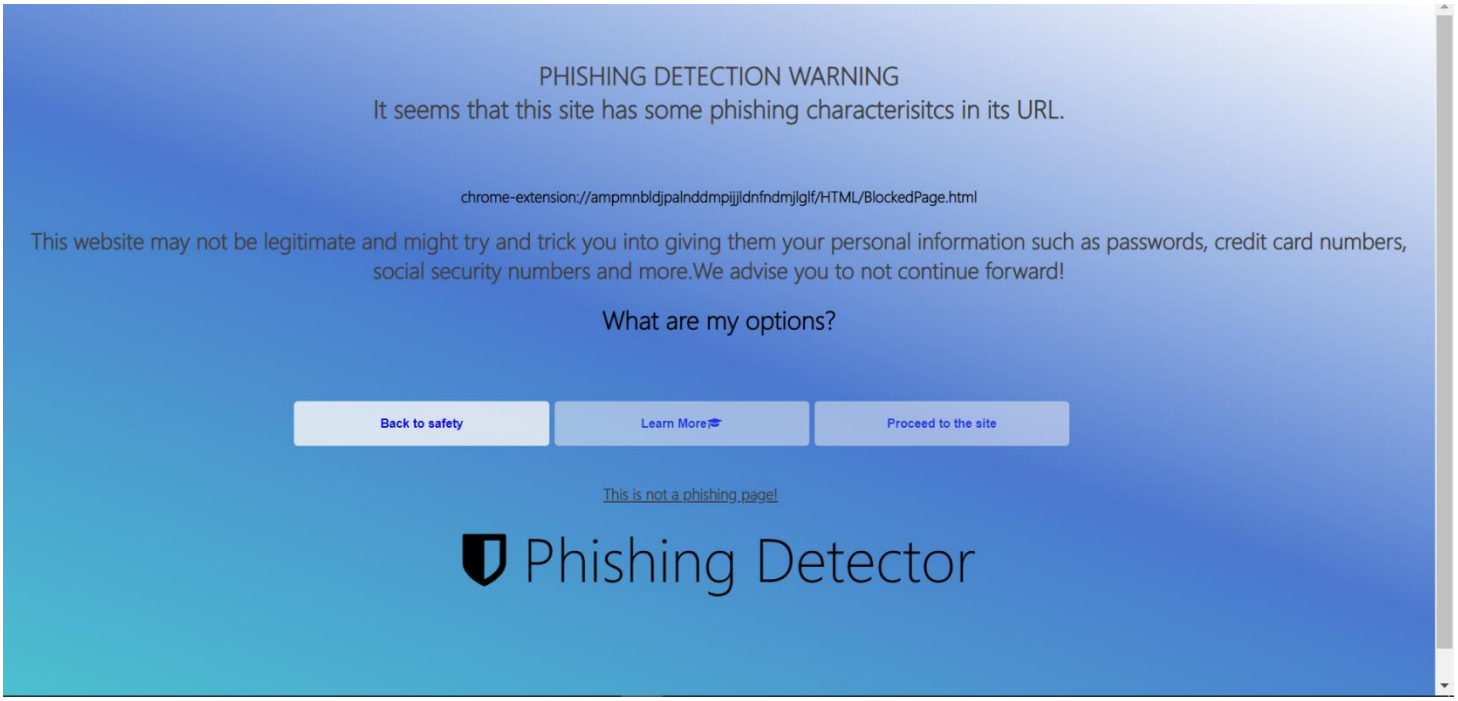


Figure 10

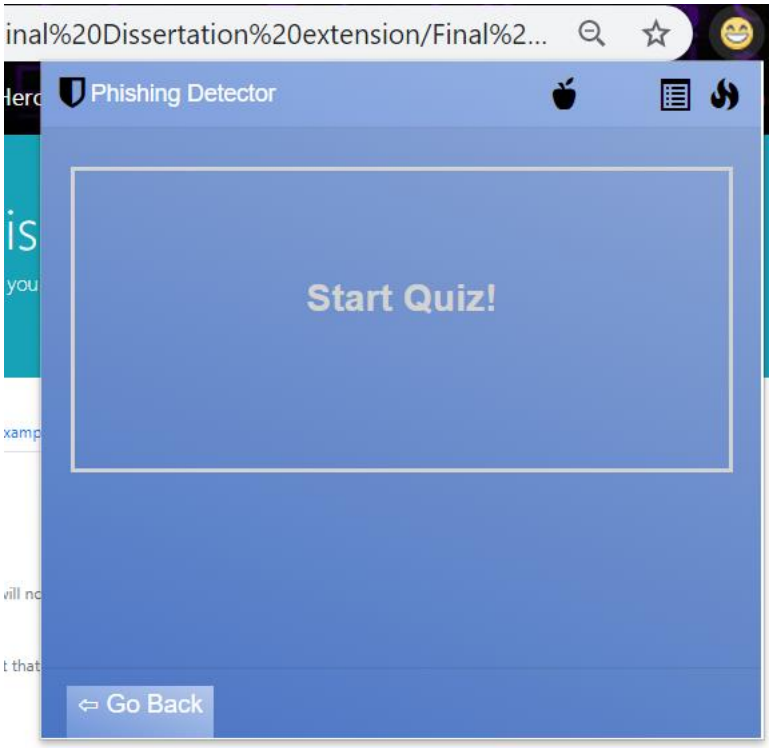


Figure 11

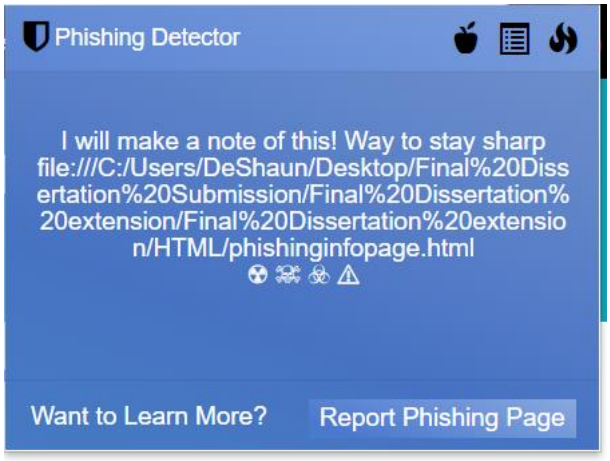


Figure 12

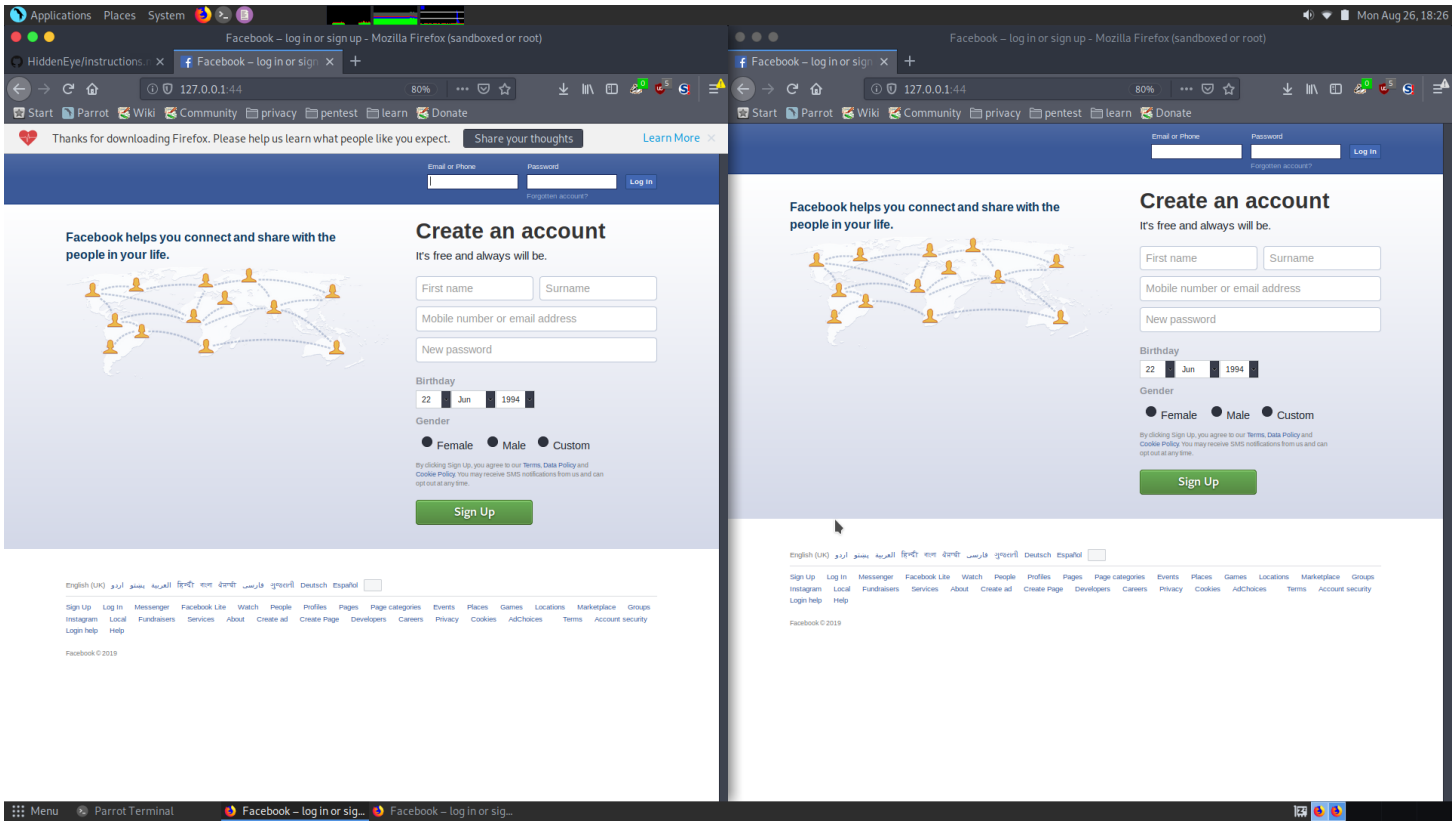
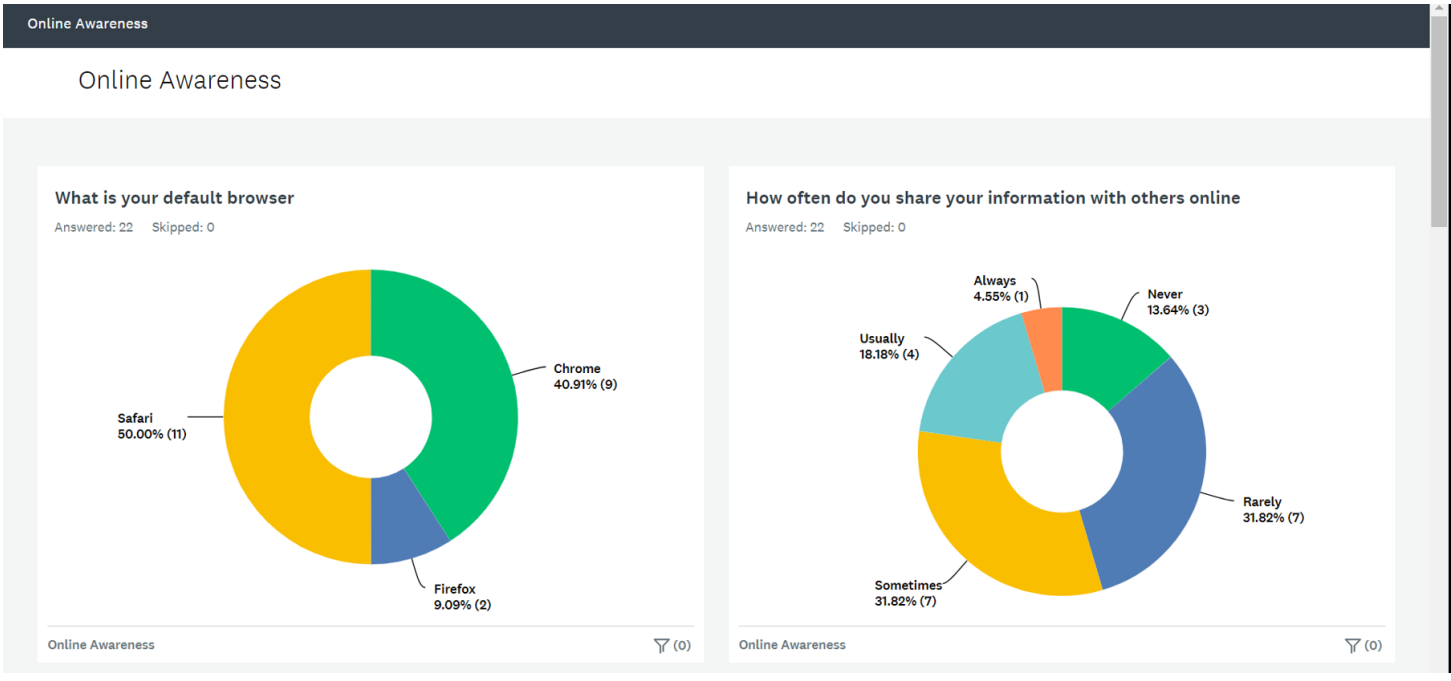
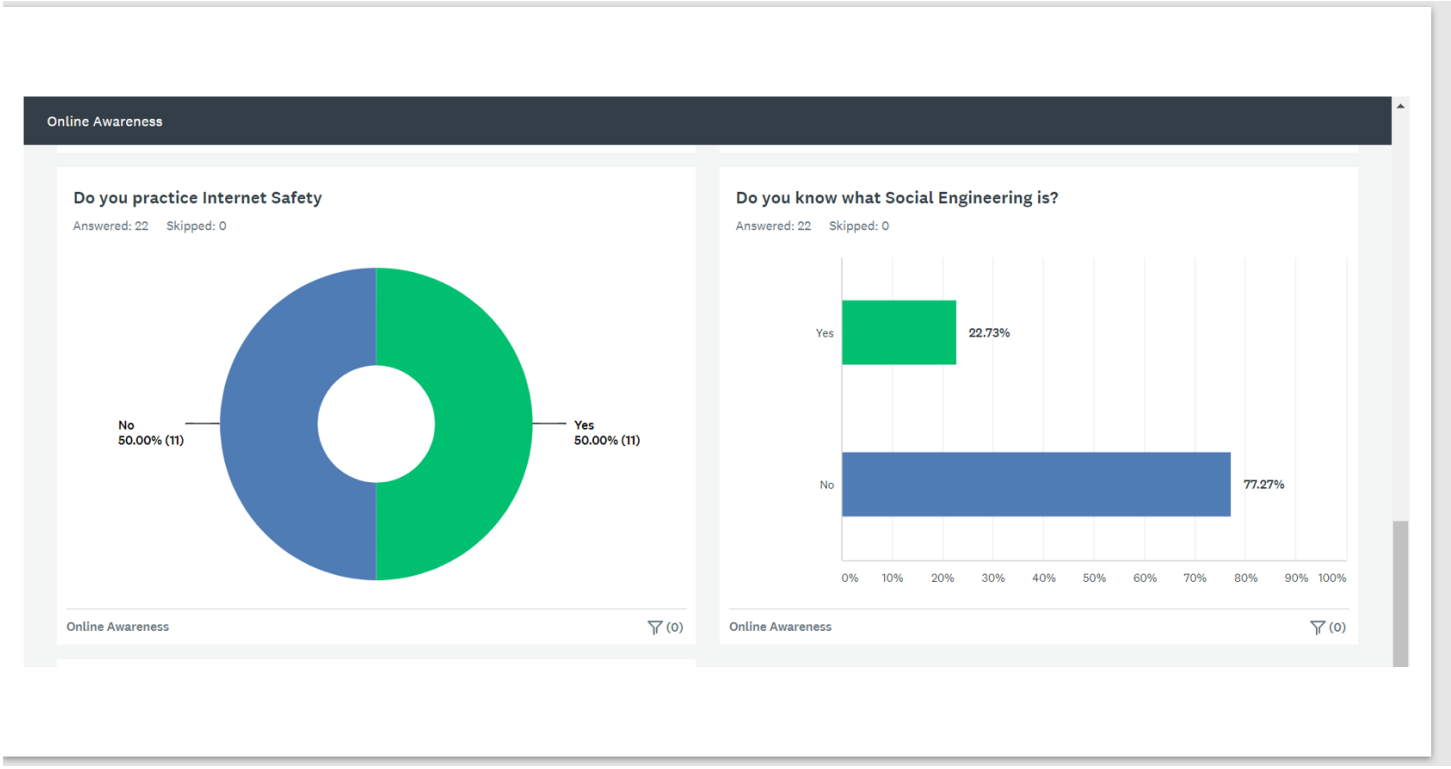
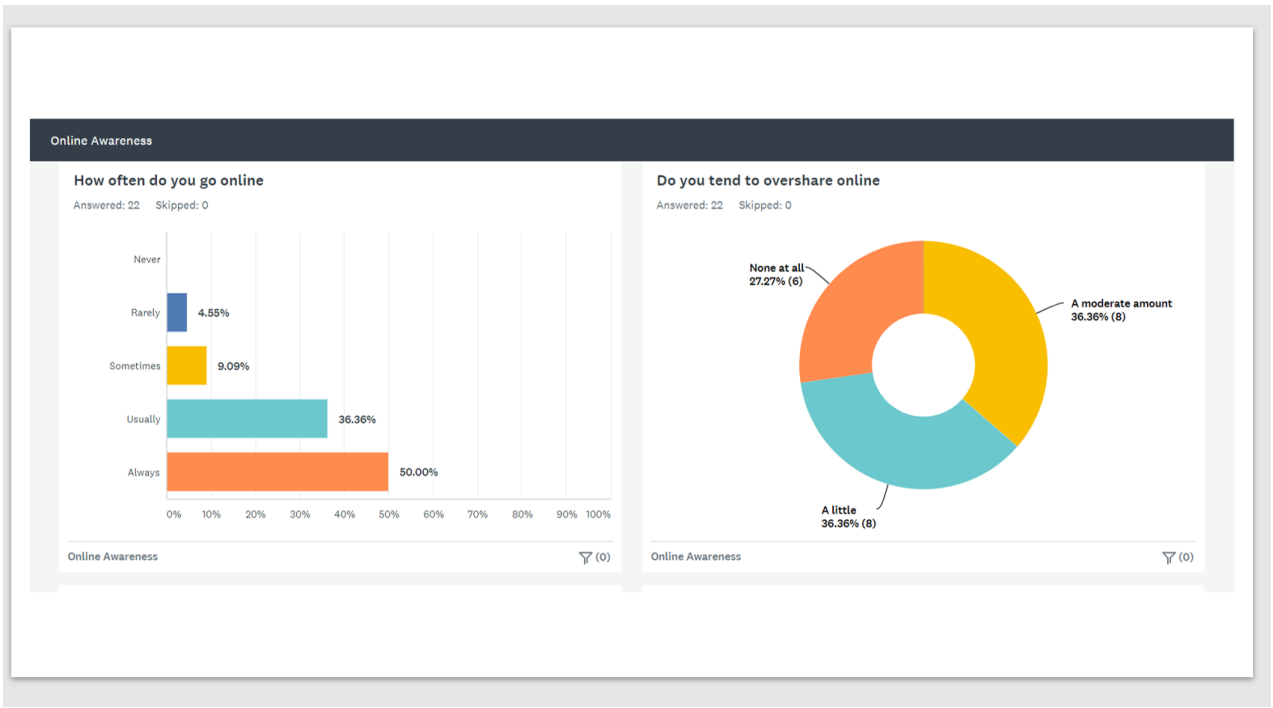

Applications Places System
Mon Aug 26, 18:26
Facebook – log in or sign up - Mozilla Firefox (sandboxed or root)
127.0.0.1:44
Start Parrot Wiki Community privacy pentest learn Donate
Thanks for downloading Firefox. Please help us learn what people like you expect.
Share your thoughts
Learn More
Email or Phone
Password
Log In
Forgotten account?
Facebook helps you connect and share with the people in your life.
Create an account
It's free and always will be.
First name
Surname
Mobile number or email address
New password
Birthday
Gender
Female
Male
Custom
Sign Up
Menu
Parrot Terminal

Figure 13


Online Awareness
Do you hover over links before visiting them
Answered: 22 Skipped: 0
No
36.36% (8)
Yes
63.64% (14)
Online Awareness
(0)
How often do you check if the website matches the URL
Answered: 22 Skipped: 0
Always
9.09% (2)
Never
31.82% (7)
Usually
31.82% (7)
Rarely
9.09% (2)
Sometimes
18.18% (4)
Online Awareness
(0)


Online Awareness
Online Awareness
What is your default browser
Answered: 22 Skipped: 0
Chrome
40.91% (9)
Safari
50.00% (11)
Firefox
9.09% (2)
Online Awareness
(0)
How often do you share your information with others online
Answered: 22 Skipped: 0
Always
4.55% (1)
Never
13.64% (3)
Usually
18.18% (4)
Rarely
31.82% (7)
Sometimes
31.82% (7)
Online Awareness
(0)


Online Awareness
Do you practice Internet Safety
Answered: 22 Skipped: 0
No
50.00% (11)
Yes
50.00% (11)
Online Awareness
(0)
Do you know what Social Engineering is?
Answered: 22 Skipped: 0
Yes
22.73%
No
77.27%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Online Awareness
(0)

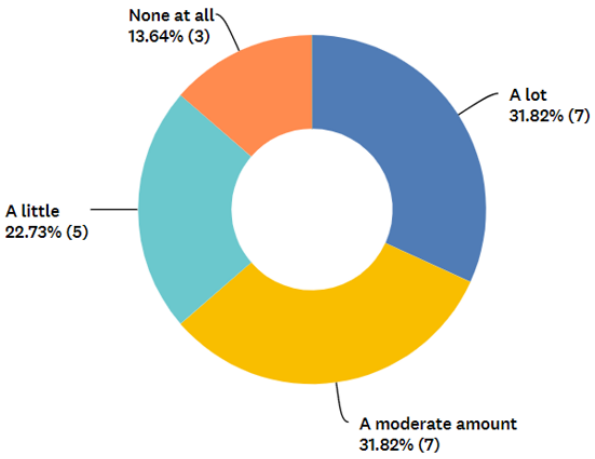

Online Awareness
How often do you go online
Answered: 22 Skipped: 0
Never
Rarely
4.55%
Sometimes
9.09%
Usually
36.36%
Always
50.00%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Online Awareness
(0)
Do you tend to overshare online
Answered: 22 Skipped: 0
None at all
27.27% (6)
A moderate amount
36.36% (8)
A little
36.36% (8)
Online Awareness
(0)

Online Awareness

Online Awareness (0)

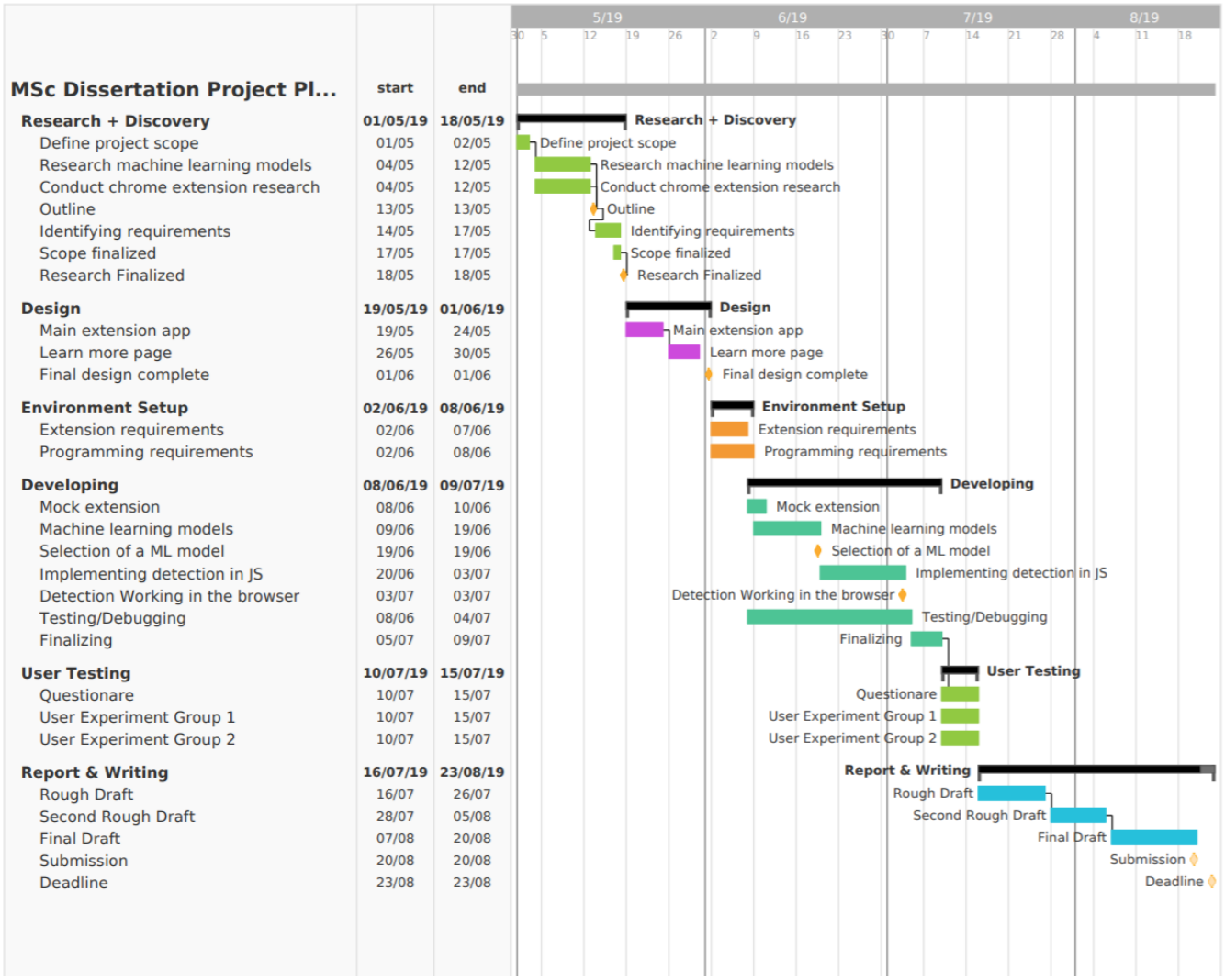
Do you know what Phishing is.

Answered: 22 Skipped: 0



Online Awareness (0)

Figure 14


MSc Dissertation Project Pl...
start
end
5/19
6/19
7/19
8/19
30 5 12 19 26 2 9 16 23 30 7 14 21 28 4 11 18
Research + Discovery 01/05/19 18/05/19
Define project scope 01/05 02/05
Research machine learning models 04/05 12/05
Conduct chrome extension research 04/05 12/05
Outline 13/05 13/05
Identifying requirements 14/05 17/05
Scope finalized 17/05 17/05
Research Finalized 18/05 18/05
Design 19/05/19 01/06/19
Main extension app 19/05 24/05
Learn more page 26/05 30/05
Final design complete 01/06 01/06
Environment Setup 02/06/19 08/06/19
Extension requirements 02/06 07/06
Programming requirements 02/06 08/06
Developing 08/06/19 09/07/19
Mock extension 08/06 10/06
Machine learning models 09/06 19/06
Selection of a ML model 19/06 19/06
Implementing detection in JS 20/06 03/07
Detection Working in the browser 03/07 03/07
Testing/Debugging 08/06 04/07
Finalizing 05/07 09/07
User Testing 10/07/19 15/07/19
Questionare 10/07 15/07
User Experiment Group 1 10/07 15/07
User Experiment Group 2 10/07 15/07
Report & Writing 16/07/19 23/08/19
Rough Draft 16/07 26/07
Second Rough Draft 28/07 05/08
Final Draft 07/08 20/08
Submission 20/08 20/08
Deadline 23/08 23/08