PhishLens

Detection of Phishing attacks

Kajal Agarwal, Sai Supraja Malla, Rudy Wahjudi,
Prateek Sharma
Computer Software Engineering Department,
San Jose State University, San Jose, CA, USA

***Abstract*—Phishing is a process to steal user's sensitive information over the internet by tricking a user to enter his information on a disguised or a fake site which is a replica of the authentic site. Leading web portals get trapped because of phishing attacks each day. Thousands of passwords are stolen without either the users or the portals knowing about it, sometimes users also lose money from their bank accounts. Companies end up paying a lot of money in the courts to deal with such attacks.**

***Keywords* – Phishing, web security, cyber security, attacks, chrome, malicious, machine learning.**

I. Introduction

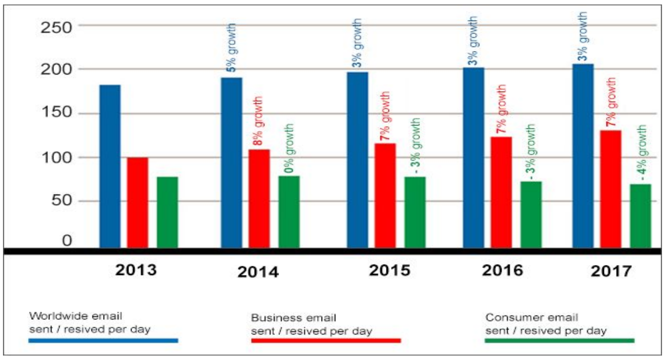
In this project, we aim to solve the problem of phishing faced by users across the world, using machine learning to eliminate the risk of phishing attacks. And therefore, providing an extra layer of security to the users.

We identify that Google Chrome browser is a popular web browser and a Chrome extension is an easier way for users to use our product and services. For this reason, we are planning to build a chrome extension with machine learning capability to detect phishing sites and notify the user in real time.

II. Phishing Statistics

While the % of development in the customer and business email accounts is not undergoing a change with each year, it is very likely that the business email movement will show development year on year (more business executed on email), while buyer email activity will decrease (perhaps because of online networking).

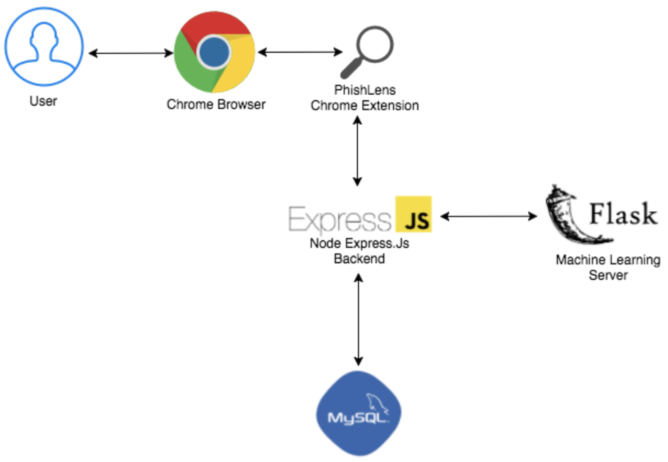
The year wise statistics[4] for phishing associated with business email, worldwide email and consumer email and their respective growth is depicted below :



III. Classification Of Phishing Attacks

On a more extensive point of view, phishing assaults can be grouped into two classifications: social building or misleading phishing and malware-based phishing assaults. Social building phishing assaults by and large draw in the mental misuse of clients or trap the organization workers into giving over their private information[3]. These assaults happen through phone messages, which appears to be genuine generally or some other social stages that interests to specific feelings in the causality, where causality winds up in click a vindictive connection or discharging touchy data, The clients with less specialized aptitude fell effectively for social building assaults, so tries must set endeavors to instruct representatives against these assaults, so as to remain two stages in front of programmers and keep these assaults from succeeding. Similarly, malware-based phishing draws in running malevolent programming or superfluous projects on the client's machine. This is a general risk for little and medium organizations. Assistance to these assaults can be delegated: keyloggers/screen lumberjacks, Man-in-the-Middle Phishing, session capturing, have document harming, DNS phishing, Search Engine Phishing and substance infusion. A portion of the techniques or measures used to bear on these phishing assaults is outlined in the accompanying sections.

IV. Architecture



Modules	Technologies used
Chrome Extension	HTML5, CSS, JQuery, Bootstrap, AJAX
Backend Server	Node.js, Express.js, MySQL, Amazon EC2
Machine Learning Server	Python, Flask, SciKit Learn, Pandas, Amazon EC2

V. Implementation

Chrome Extension is the client that is interfacing with the user on the Google Chrome browser. Every time user enters a new URL on the address bar, PhishLens will make an API call to the backend server. Backend server will send back a response to PhisLens. Correspondingly, PhishLens will notify a user if the URL is a phishing site.

Backend server receives a request from PhishLens Chrome extension to check whether the URL is pointing to a genuine site. Then, backend server will make an API call to the Machine Learning server to check the validity of the URL.

Machine Learning server uses a decision tree classifier to predict whether a URL is a phishing site or not. The dataset that is used to train this classifier was prepared from the list of URLs (Phishing URLs and legitimate URLs). We converted each of the URLs to a set of parameters that corresponds to the URL rules output. Prior to running through these rules, we will also check with PhishTank API whether the URL is already reported as a phishing site in their database or not. If it is not in the PhishTank database, URL will be converted into parameters and then fed into decision tree classifier.

VI. Machine Learning Rules

These are the rules that we have implemented to identify URL characteristics [1].

Rule 1: Check for URL with IP Address

-1	URL does not contain an IP address.
1	URL contains an IP address.

Rule 2: Check for URL length

-1	URL length < 54
0	54 <= URL length <= 75
1	URL length > 75

Rule 3: Check for URL in shortening format

-1	URL is not using shortening format
1	URL is using shortening format

Rule 4: Check for URL that has "@" symbol

-1	URL does not have "@" symbol
1	URL has @ symbol

Rule 5: Check for URL that has "//" redirection

-1	URL does not have "//" redirection
1	URL has "//" redirection

Rule 6: Check for URL that has "-" symbol

-1	URL does not have "-" symbol
1	URL has "-" symbol

Rule 7: Check for numbers of Sub Domain or Multi Sub-domains in the URL

We check this by removing the valid country-code top-level domains (ccTLD) and second-level domain(SLD) and then count the remaining dots.

-1	URL only has one dot.
0	URL has two dots
1	URL has more than two dots

Rule 8: Check for favicon validity

-1	Favicon is loaded from the same domain
1	Favicon is missing or loaded from different domain

Rule 9: Check if the site server using non-standard port

-1	Site server opens all of the non-standard port
1	Site server blocks all port

Rule 10: Check if URL has the word "HTTPS" in the domain name

-1	URL doesn't have the word "HTTPS" in its domain
1	URL has the word "HTTPS" in its domain

For example:
https://account-security-system.cf/recovery-chekpoint-login.html

If we ran through the above site to the Machine Learning Rules, the resultant parameters will look something like below:

[-1,0,-1,-1,-1,1,-1,-1,1,-1,1]

These parameters are then fed to the decision tree. The decision tree internally compares these parameters with the existing set of datasets and returns a boolean valueof True if URL is a phishing site and False otherwise.

VII. Conclusion

In this project, we have built a Chrome Extension that has successfully detected phishing URLs. Our implementation is not 100% accurate, however, we can increase detection accuracy by doing the following:

1. Train the Machine Learning classifier with larger dataset so that it can improve the prediction accuracy of detecting phishing sites.

2. Implement additional machine learning rules to differentiate a common pattern shared among phish site and non-phish site.

VIII. Acknowledgment

We would like to thank Professor Rakesh Ranjan, Dept. of Computer Software Engineering, San Jose State University, for immense motivation and guidance throughout the project.

IX. References

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Nicolas Papernot (2016) Detecting phishing websites using a decision tree [https://github.com/npapernot/phishing-detection]

[3] Sheng S, Magnien B, Kumaraguru P, Acquisti A, Cranor LF, Hong J, Nunge E (2007) Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In: Proceedings of the SOUPS, Pittsburg, pp 88–99

[4] Sara Radicati, PhD; Principal Analyst: Justin Levenstein. Statistics Report 2013-2017.The Radicati Group, Inc. A Technology Market Research Firm.