Malicious URL Detection Using Machine Learning

Patil Bhavesh , Patil Pranjal , Pisal Omkar , Tupe Ketan , Dr. P. B. Kumbharkar

[1, 2, 3, 4](Students, Computer Engineering, JSPMs Rajarshi Shahu College of Engineering Pune)

[5](Professor, Computer Engineering, JSPMs Rajarshi Shahu College of Engineering Pune)

Abstract: *Phishing is a form of cybercrime where an attacker imitates a real person / institution by promoting them as an official person or entity through e-mail or other communication mediums. In this type of cyber-attack, the attacker sends malicious links or attachments through phishing e-mails that can perform various functions, including capturing the login credentials or account information of the victim. The standard way to specify page layouts is through the style sheet (CSS), the developed algorithm detects similarities in key elements related to CSS. Phishing detection includes approach that uses profiles of trusted websites' appearances to detect phishing.*

*Keywords***:** *t*Mobile phones; phishing attack; security; anti-phishing

I. INTRODUCTION

Phishing is defined as the fraudulent acquisition of confidential data by the intended recipients and the misuse of such data. **The phishing attack is often by email.** An example of Phishing; as if e-mail app be from known web sites, from a user's bank, credi company, e-mail, or Internet service provider. Gene personal information such as credit card numb password is asked to update accounts.

These emails contain a URL link that directs us another website. This site is actually a fake or mo website. When users go to this site, they are asked to personal information to be forwarded to the phi attacker.

does not access the content of messages and theref privacy-preserving .

.

Literature Survey

Neda Abdelhamid, Fadi Thabtah, Hussein Abdel-jaber "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features" IEEE 2017.

Experimentally compare large numbers of machine learning techniques on real phishing datasets and with respect to different metrics. The comparison is to reveal the advantage and disadvantages of machine learning predictive models and to show their actual performance when it comes to phishing attacks.

For comparison of suspicious websites with image database SURF is used.

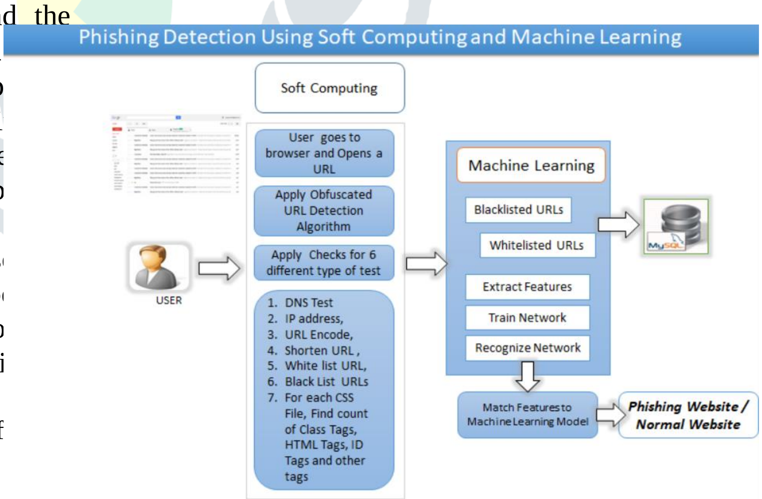
2. Tianrui Peng, Ian G. Harris, Yuki Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning", in proceedings of the 12th IEEE International Conference on Semantic Computing, IEEE 2018.

An approach which uses natural language processing techniques to analyze text and detect inappropriate statements which are indicative of phishing attacks.

3. Muhammet Baykara, Zahit Ziha Gurel, "Detection of Phishing Attacks", 2018 International Conference on Collaboration Technologies and Systems, IEEE 2018.

This software, phishing and spam mails are detected by examining mail contents. Classification of spam words added to the database by Bayesian algorithm is provided.

II. PROPOSED SYSTEM


Phishing Detection Using Soft Computing and Machine Learning
Soft Computing
User goes to browser and Opens a URL
Apply Obfuscated URL Detection Algorithm
Apply Checks for 6 different type of test
1. DNS Test
2. IP address,
3. URL Encode,
4. Shorten URL ,
5. White list URL,
6. Black List URLs
7. For each CSS File, Find count of Class Tags, HTML Tags, ID Tags and other tags
USER
Machine Learning
Blacklisted URLs
Whitelisted URLs
Extract Features
Train Network
Recognize Network
MySQL
Match Features to Machine Learning Model
Phishing Website / Normal Website

The **aim** is to steal sensitive data such as credit card and login information or to install malicious software on the victim's machine. Phishing is a common type of cyber-attack that everyone must learn to protect them. Phishing is start with a fake e-mail or other type of transmission designed to attract a victim. In this type of attack, the message appears to come from a trusted source.

In a phishing attack, attackers can use social engineering and other public information resources, including social networks like LinkedIn, Facebook and Twitter, to gather background information about the victim's personal and work history, interests and activities. With this pre-discovery, attackers can identify potential victims' names, job titles and email addresses, information about the names of key employees in their colleagues and organizations.

Phishing is also used to learn someone's password or credit card information. With the help of e-mail prepared as if coming from a bank or official institution, computer users are directed to fake sites.

The common information that is stolen by a phishing attack is listed as follows:

- User account number
- User passwords and user name
- Credit card information
- Internet banking information

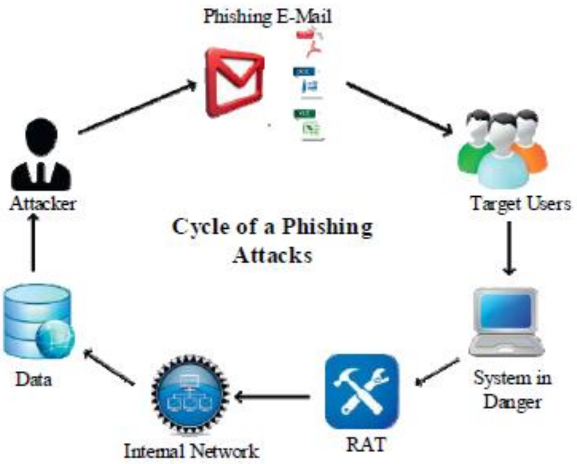


Figure: - The processing cycle of phishing attacks [1].

AVOIDING PHISHING ATTACKS

A whitelist in the context of phishing detection is simply a list of trusted websites. For CSS detection to work properly, the list contains more than just the URL of the trusted website. Each entry in the whitelist database contains six strings: the URL of the trusted site, the domain of the site, the title of the site, the CSS filename, the CSS domain, and the CSS content of the file.

a. The URL of the trusted site:

The URL of the trusted site is used to periodically update the CSS information in the database. This is the URL of the site such as “https:\\signin.ebay.com”.

b. The domain of the site:

The domain of the trusted site is the domain of the URL such as “signin.ebay.com” and is used to determine whether the current page displayed in the browser is on the whitelist or not.

c. The CSS filename:

The CSS filename is the filename of the CSS file such as “paypal.css” and can also be used during CSS content detection to speed up detection by matching potential phishing site CSS filenames with filenames in the whitelist database.

d. The CSS domain:

The CSS domain is the domain of the location of the CSS file such as “secureinclude.ebaystatic.com”. Often the domain is the same as the site domain, but in other cases such as eBay, the CSS file is hosted on a different domain. Storing the CSS domain is essential because if a match is found on a website not in the whitelist, then it is most likely a phishing site linking to the actual CSS file location of the legitimate site.

e. The CSS content of the file:

The CSS content is the actual text contained in the CSS file that contains all of the style information. The CSS content is used to compare with the CSS content of a possible phishing site in order to determine if there is a match with a legitimate site.

III. MATHEMATICAL MODEL

Let us consider S be a Systems such that

S= U, ES, SS, K, DE, DS, where

ˆ U= {U1, U2, U3.Un | U is a Set of all USERS }

U is the users of the system. Users of the system may grow as the system is used by more and
more people. User is infinite set.

ˆ ES =ES1,ES2 | ES is a Set of user visit to the browser and opens the a URL} These are the data to
be entered in URL of the system, so this is also Finite Set.

ˆ SS= {SS1,SS2,SS3, SSn | SS is a Set of features checked for detection}

SS are the main features like DNS Test, IP address, URL encode, Shorten URL, White List and Black
List URL so this is also Finite Set.

ˆ K= {K1, K2, K3.Kn | K is a Set of train network}

This set is used for training the network. This is also an infinite Set.

ˆ B= {B1, B2, B3.Bn | Bn is a set for Recognize Network }

Event 1

User will visit the browser and opens the URL. Let f(U) be the function of user. Thus f(U) -> {Ds U Ss}

Event 2

Features checked for detection. Let f(U) be a function of user. Thus f(U)-> {Es U Ss}

Event 3

Feature understanding.

Let f(Es) be a function to understand main features used URL detection.

Event 4

Check for different types of test.

Let f(Es) be a function of courting the class tags, hash tags, ID tags and other tags.

Thus f(Es)-> {F1,F2,F3,…,Fn}.

Event 5

Training and recognize network. Let f(Un) be a function of system. Thus f(Un)-> Es

Event 6

Matching features to the machine learning model. Let f(Es) be a function of test. Thus f(Un)->Es

IV. CONCLUSION

In this survey paper, we have surveyed some technique of different researchers that was applied to detect phishing sites effectively. The phisher have to create a phishing website to lure the victim which seems as legitimate one. In the proposed technique, the system model is built to detect phishing sites by using some algorithms like CSS Detection Algorithm, ObURL Detection Algorithm and Feed Forward Neural Network.

.

REFERENCES

[1] Neda Abdelhamid, Fadi Thabtah, Hussein Abdel-jaber “Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features” IEEE 2017.

[2] Tianrui Peng, Ian G. Harris, Yuki Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning”, in proceedings of the 12th IEEE International Conference on Semantic Computing, IEEE 2018.

[3] Muhammet Baykara, Zahit Ziha Gurel, “Detection of Phishing Attacks”, 2018 International Conference on Collaboration Technologies and Systems, IEEE 2018.

[4] Neda Abdelhamid, Fadi Tabtah, Hussein Abdel-Jaber,”Phishing Detection : Recent Intelligent Machine Learning comparison based on models content and features”, IEEE 2017.

[5] Frank Vanhoenshoven, Gonzalo Napoles, Rafael Falcon, Koen Vanhoof and Mario Koppen,”Detecting Malicious URLs using Machine Learning Techniques” on Universiteit Hasselt Campus Diepenbeek Agroralaan Gebouw D, BE3590 Diepenbeek, Belgium, IEEE 2016.

[6] Guang-Gang Geng, Zhi-Wei Yan, Yu Zeng and Xiao-Bo Jin “RRPhish- Anti-Phishing via Mining Brand Resources Request” 2018 IEEE International Conference on Consumer Electronics (ICCE)

[7] Muhammet Baykara and Zahit Ziya Gürel “Detection of phishing attacks” IEEE 2018