A Phishing Detection System for E-Banking Site

Charudutt D. Pandya[1] Director Dev M. Rathod[2]
[1]PG Student [2]director
[1]Department of computer engineering
[1]Gujarat Technological University. Ahmedabad [2]Prediqnous Cyber Security and I.T Intelligence, Ahmedabad

Abstract— Phishing is an attack that deals with social engineering methodology to illegally acquire and use someone else's data on behalf of legitimate website for own benefit (e.g. Steal of user's password and credit card details during online communication). It is affecting all the major sectors of industry day by day with a lot of misuse of user credentials. It is affecting all the major sectors of industry and banking day by day with a lot of misuse of user credentials. To protect bank's users against online phishing, various anti-phishing techniques have been proposed that follows different strategies like client side and server side protection. Hence it is necessary to detect the attacker and secure individuals private data. Till now many approaches are found to detect phishing websites. Among them are Bacterial Foraging algorithm, Visual similarity based approach, etc. Every approach has some weakness or limitations such as less efficient or time consuming or no up-to-date blacklist or phishtank. To detect the phishing websites different techniques are proposed such as classifier technique, heuristics based technique, hybrid technique etc. In this paper, we propose a technique based on Heuristic based technique. In first Phase, website is identified based on different parameters like URL, IP Address, Forms, age of domain. In second Phase, the visual similarity of the phishing banking page is compared with the original banking website. The proposed approach also gives the suggestions to the users for their particular domain search that makes user more comfortable to use the system.

Key words: URL, IP Address, Forms, age of domain,E-Banking.

I. INTRODUCTION

Under the domain of computer security, Phishing is the illegally deceitful process of trying to acquire confidential information just as usernames, passwords and credit card details, by impersonate as a legitimate thing in broadcasting. Phishing is a deceitful e-mail that tries to take you to divulge secret data that can then be used for illegitimate purposes .There are different types of this scheme. It is feasible to theft identity for confidential information in supplement to usernames and passwords just as credit card numbers, bank account numbers, social confidential numbers and mother's maiden names. Phishing presents direct threats through the use of stolen credentials and secondary threat to institutions that conduct business on line through erosion of customer confidence. The damage generated by Phishing ranges from contradiction of access to email to substantial financial loss.

In state for Internet thefts to purposefully "phish" your secret data, they sends an email to a website. Phishing emails will encourage you to click on a link that shifts you to a site where your sensitive information is requested. Trustworthy organizations would never request this information of you via email.

In general, phishing attacks are performed with the following four steps:

(1) A fake web site which looks exactly like the legitimate Web site is set up by phisher

(2) Phisher then send link to the fake web site in large amount of spoofed e-mails to target users in the name of legitimate companies and organizations, trying to convince the hypothetic victims to visit their web sites.

(3) Victims visit the fake web site by clicking on the link and input its useful information there.

(4) Phishers then steal the personal information and perform their fraud such as transferring money from the victims' account. Figure 1 depicts the process of phishing.[2]

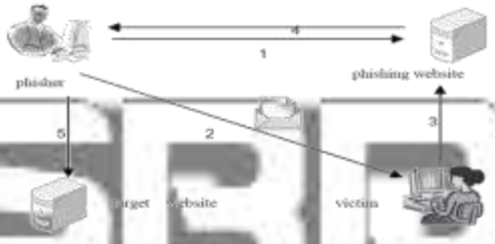

Fig. 1: Process of Phishing

There are thousands of fake phishing websites established online every day, luring a number of clients. According to a phishing activity trend report published by Anti-phishing working group on 7 April 2014, a lot of phishing attacks were done in fourth half of year 2013 as can be seen from Fig 2. The number of unique phishing reports submitted to APWG[2].

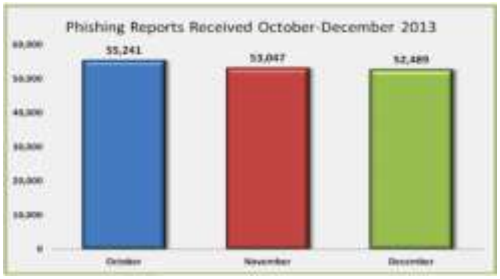

Fig. 2: Phishing websites report[1]

Financial Services continued to be the most targeted industry sector throughout 2013. Payment Services has experienced the most drastic changes during 2013, little decreasing from 56.30 percent in 3rd quarter 2013 to 53.95 percent in 4th quarter 2013[1].

Seeing financial service sector and payment service sectors deals with money transactions it can be concluded that main objective of phishers is to steal financial details of victims and misuse that for their own gain. Retail sector appears to be third most vulnerable and classified as the least vulnerable to phishing attacks. So phishing attacks are emerging as one of the major area where immediate concern is needed as it is affecting all the major sectors of industry creating a lot of loss.

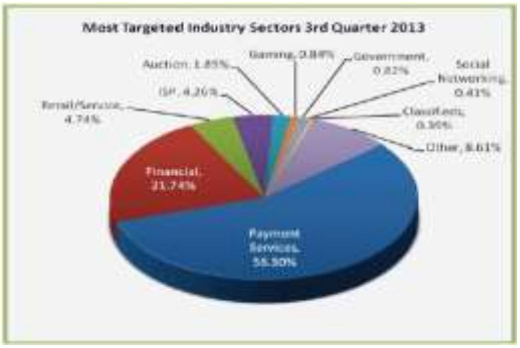

Fig. 3: Industry sector area wise affect of Phishing

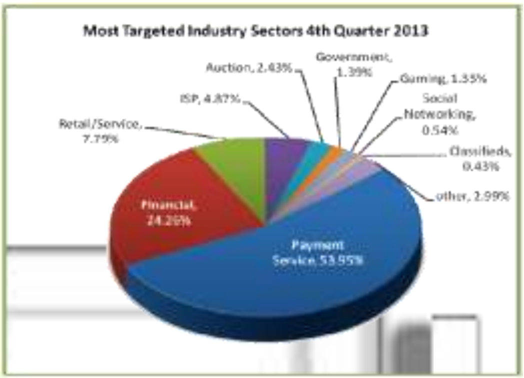

Fig. 4: Industry sector area wise affect of Phishing

Till now many approaches are proposed for detecting phishing websites such as AZProtect, Black list Generator, Bacterial Foraging algorithm, classifier algorithm etc. All these approaches use different algorithms or methods for detecting phishing websites. As per the websites characteristics or behavior we can able to detect whether it is phishy or genuine. However people fail to detect the phishing websites, because now-a-days attackers are also very clever. It is necessary to propose the best technique to detect the phishing websites.

II. LITERATURE REVIEW

Phishing attacks create serious problem to the e-banking and the e-commerce websites. Through these websites attacker takes the user name, passwords and the account number of the customer and fetch the privileges of the customer. Both consumer and the financial organizations are at threat for huge amount of fake transactions via stolen data. The treat is rapidly increasing, the victims being clients or users of financial or banking organisations, trading corporations, and supplier of internet services. In spite of lot of work that has been done on implementing better and efficient tools on phishing detection and prevention, still it is very hard to completely eradicate the problem and to estimate no. of users that actually caught in bait of phishing as victim.

There are a lot of indicators that identifies and distinguish legitimate sites from phishing sites. Based on case studies conducted 27 features and indicators were gathered and clustered them into six criteria. Those six criteria are URL & domain identity, Security & encryption, Source code & java script, Web address bar, Page style & contents, and Social human factor[2].

Criteria	Phishing Indicators
URL & Domain Identity	Using IP address
	Abnormal Request URL
	Abnormal URL of Anchor
	Abnormal DNS record
	Abnormal URL
Security & Encryption	Using SSL Certificate
	Certificate Authority
	Abnormal Cookie
	Distinguished Names Certificate
Source Code & Java Script	Redirect Pages
	Straddling Attack
	Pharming Attack
	OnMouseOver to hide the link
	Server from Handler
Page style & Contents	Spelling Errors
	Copying Website
	Using forms with Submit Button
	Using pop-up windows
	Disabling Right Click
Web Address Bar	Long URL Address
	Replacing similar char for URL
	Adding a prefix or suffix
	Using the @ symbol to confuse
	Using the hexadecimal char codes
Social Human Factor	Emphasis on security
	Public generation salutation
	Buying time to access accounts

Table 1: Phishing indicators with its criteria

A. *Anti-Phishing:*

Anti-phishing refers to the method employed in order to detect and prevent phishing attacks. Anti-phishing defends users from phishing. A lot of work has been done on phishing detection devising various phishing detection techniques. Some technique works on emails, some works on attribute of web sites and some on URL of the websites. Many techniques focus on enabling clients to recognize and filter various types of phishing attacks. In general phishing detection techniques can be classified into following two main categories namely User Awareness and Software detection. Software detection techniques have four sub categories which are depicted in Fig 5[4].

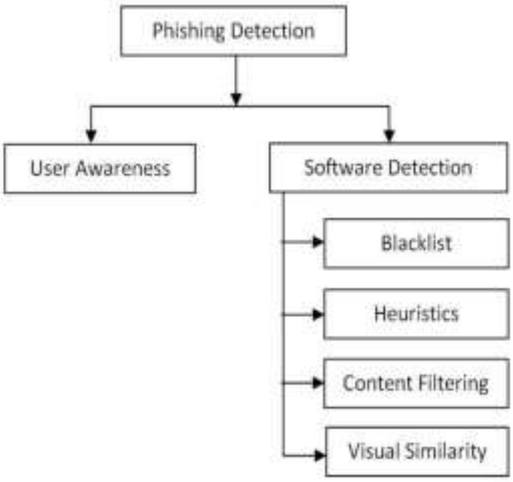

Fig. 5: A Phishing Detection Approaches

B. *User Awareness:*

End users can be educated to better understand the nature of phishing attacks, which ultimately leads them into correctly identifying phishing and non-phishing messages. This is contrary to the categorization in where user training was considered a preventative approach. However, user training approaches aim at enhancing the ability of end-users to detect phishing attacks.

C. *Software Detection:*

These mitigation approaches aim at classifying phishing and legitimate messages on behalf of the user in an attempt to bridge the gap that is left due to the human error or ignorance. This is an important gap to bridge as user-training is more expensive than automated software classifiers, and user training may not be feasible in some scenarios.

1) *Blacklist Based Approach:*

Blacklist is collection of known phishing Websites and addresses published by trusted entities like Google's and phishtank black list. It requires both a client and a server components. The client component is implemented as either an email or browser plug-in that interacts with a server component, which in this case is a public webwsite that provides a list of known phishing sites.

2) *Heuristic Based Approach:*

This technique classifies the URL's based on certain heuristics that are generally observed in phishing sites but this does not guaranteed to classify the URL's correctly but it performs better than blacklisting.

3) *Content Filtering Approach:*

In this methodology Content or email are filtered as it enters in the victim's mail box using machine learning methods, such as Bayesian Additive Regression Trees (BART) or Support Vector Machines (SVM).

4) *Visual Similarity Based Approach*

This method is used to measure the similarity between two given web pages by calculating the similarity between the content elements (text, image, layout) contained in the web pages. Algorithms are used to compute visual similarity to detect the phishing web pages which have higher similarities to phishing targets.

D. *Techniques of Phishing Website Detection:*

E. *Bacterial Foraging Algorithm:*

Classifies websites as per their characteristics as well as their content.
Limitation: Less accurate, Time consuming.

1) *Statistical Learning Theory:*

Spoofed and concocted websites are classified as their performance through SLT based classifiers.
Limitation: Difficult to identify concocted sites, Less accurate for phishing websites.

2) *Phishnet:*

Identify websites as per their content matching as well as their DNS and URL.
Limitation: No updated blacklist.

3) *Finite State Machine:*

Demonstrating behavior or responses with respect to input submissions and classifies as per different heuristics.
Limitation: Efficient for only web applications, Not reported suspected websites directly

4) *Visual Similarity:*

Detect only exact same fake web pages by comparing images with registered database.
Limitation: Lack of prior knowledge about priori knowledge about web pages.

5) *WHOIS feature:*

Classifiers classify as per website's URL and content features.
Limitation: Lack of regularity content, No updated blacklist.

6) *Blacklist Generator:*

Generating blacklist as per Google's Top-10 search or Phishtank Phishing website dataset and creating blacklist.
Limitation: Not accurate for recent websites, cannot get exact search.

7) *Hierarchical clustering:*

automatic phishing categorization by extracting different features of websites
Limitation: Less efficient, Not accurate as well others.

8) *CANTINA+:*

Filters phish sites using hash value as well as login form filtering.
Limitation: More computation needed, more time consuming, Attackers can compromise legitimate domains.

9) *PageRank Based:*

Classifies phishing websites as per heuristics and Google's Top-10 searches.
Limitation: More heuristics, Calculation is complex.

III. PROPOSE WORK

A. *Problem Formulation:*

Phishing attacks are based on identity theft of genuine banking website or web application. Phishing attack is going through below strategy. As per figure, for phishing attack attacker visits genuine banking website and extract all the information about website. As per all information about webpages attacker creates same replica of that website for attracting the user. After creating phishing banking website attacker send the links of phishy website to clients through email, sms, instant messengers, etc. Then attacker attacks to the end users and asks for confidential information about the user. Attackers gain all these personal information for misuse or for gaining privileges from the users.

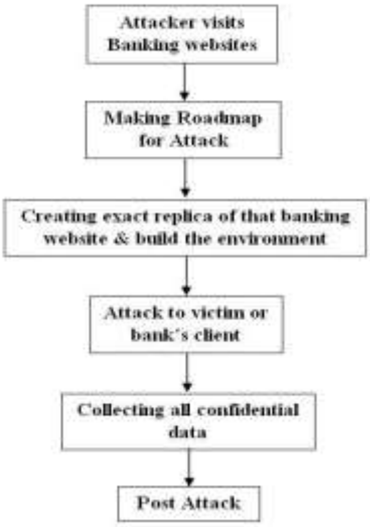
Attacker visits Banking websites
Making Roadmap for Attack
Creating exact replica of that banking website & build the environment
Attack to victim or bank's client
Collecting all confidential data
Post Attack

Fig. 6: Strategy of Attack

There are three techniques to classify phishing websites.

1) *Classifier Technique:*

in which blacklist is created and the heuristics are measured for classifying the phishing webpages.

2) *b) Lookup System:*

in which blacklist is created using IE Phishing filter and creating whitelist, too.

3) *c) Hybrid System*

in which combination of both techniques which are creating blacklist as well as whitelist and calculating heuristics.

Phishing solutions can be classified in below categories:

4) *Blacklisting:*

In this solution, comparing URL with the blacklist. If the URL matches with the blacklist then alert the user for threat.

5) *Machine Learning:*

It is for about creating white list as well as blacklist and giving the result about the fake website. It gives 100% true positive but cannot control false positive.

6) *Heuristics:*

In this approach classifies the URL"s based heuristics and observing phishing sites but it is not giving guaranteed result for phishing sites like blacklisting.

7) *Trusted Communication:*

This technique is for authenticate the site for secure browsing. Hybrid: In this technique multiple features are combined for phishing site detection.

As shown in Fig 7 for visual similarity based phishing attack, attacker subscribes the genuine website and makes the duplicate of that website. Though this fake website attacker sends e-mail to victim user. Here, victim user does not know about this phishing site and filling up confidential data like account number, user id, password and birth date to this phishing site. Now attacker can use this confidential information for his/her personal use. Hence, the private data of the user can be available for the attacker website, extracting the domain name from URL and giving suggestions from Google Top-10 searches.

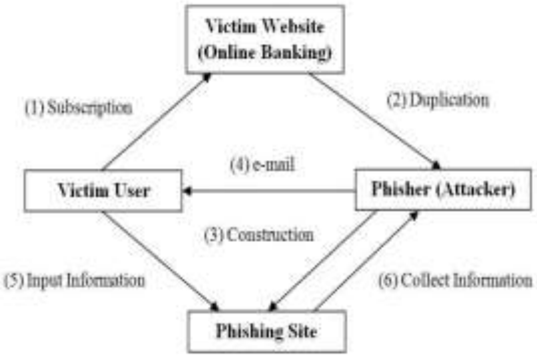

Fig. 7: Phishing Attack

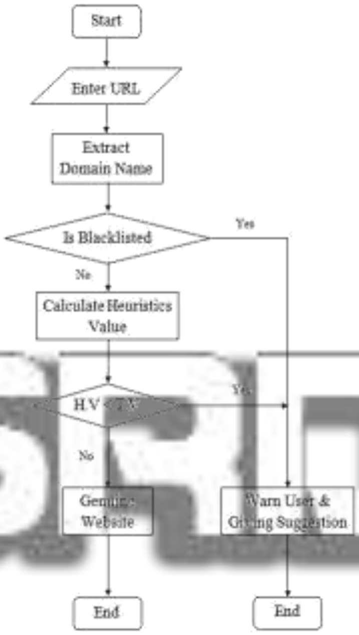

Fig. 8: Flow Diagram

User will enter the URL for identification of the website whether it is legitimate or not. Entered URL is matched with the blacklist. If the URL is within the blacklist, user will be warned for phishing site and giving suggestions for that particular search. If the URL is not matched with the blacklist, calculating heuristic values. After this, obtain the values of the heuristics. The values of heuristics values of GTR and age of domain are obtained by parsing the pages which will give these values and the values of suspicious URL and IP address are obtained by checking the URL.

8) *Bad Forms:*

In this heuristic checking for the form actions and how many links are within the particular form. Genuine websites have links which are similar to their home page or domain name.

9) *Pop-ups:*

Generally no more pop-ups into legitimate websites. Here, we have considered pop-ups no more than five.

10) *Suspicious URL:*

For heuristics, checking whether the URL is including '@' or '-'. However, valid site rarely uses „-„. In the URL after

'@' string will be considered before'@' string part is discarded. Heuristics will check whether these conditions are satisfied or not for phishing site. If the conditions are satisfied then the site is suspicious else declared as legitimate site.

11) IP Address:

URL contains IP address as its domain is checked by this heuristics.

12) Dots:

How much dots are within the URL will be checked by this heuristics. Normally, legitimate URL has less number of dots. Here, checking for minimum five dots within the URL. If there are more than five dots, the URL is considered as not legitimate site and calculating values as per this strategy. Classification algorithm will be applied after obtaining the heuristic values on the training dataset to obtain the weights using a simple forward linear model described below,

$$S=\Sigma f(w_i*h_i) \;.\;.\;.\;.\;.\;.\;.\;.(1)$$

Where h_i is each heuristic's result , w_i is each heuristic's weight. After this calculate the weight for each heuristics. For this, the higher the weight will be given to it.

$$W_i=(e_i/\Sigma e_i) \;.\;.\;.\;.\;.\;.\;.\;.\;.(2)$$

Where e_i is the effect of each heuristic and will be calculated as per above (2) equation.

From the score of (1), S of the URL will be calculated. If the value of S is greater than the threshold then it is legitimate site else warn the user as a phishing site. If the value is not equals to threshold or less than the threshold site will be declared as a phishing site and warn the user for that. After this, for user convince giving suggestions to the users. If all these heuristics will be satisfied by any website then the page source of the web page will be compared with original webpage of Google's Top-10 searches. If it matches with the original website, then declared that URL as legal web page, else legal webpage. For suggestions extract the domain name and through good search engine giving suggestions to the users which are safe from phishing.

IV. CONCLUSION

Today phishing is very serious attack so it is necessary to detect it. To detect phishing websites there are many techniques are found like visual similarity, blacklist generator, PageRank, Bacterial Foraging algorithm etc. All this techniques have drawbacks like more time consuming, more computation etc. All these drawbacks affect the performance of the approach and giving less efficiency. So it is necessary to propose a new approach with high efficiency. This approach is based on classification of URL‟s heuristics based on their weights. First of all, matching the URL within the blacklist and identifies whether it is legitimate or not. After this calculate the weights of heuristics of the URL and identifying the website by classification algorithm. Here, used classification algorithm is simple linear model approach. Calculated heuristics will be compared with the threshold value. If the value is greater than the threshold, the URL will be declared as legitimate else warn the user. After warn the user giving some suggestions through extracting domain name and searching domain name with the help of good search engine. In this approach, I tried to overcome all these limitations and control false positives.

V. FUTURE WORK

In this approach, we can detect the phishing websites and providing Google's top five links about the search. Here, detecting phishing website is based upon the different heuristic values and trying to get the best result. In future the technique can be proposed by removing or adding more heuristics to gain high accuracy rate to classify the phishing websites as well as legal websites. These techniques can be implied by combined other techniques and develop hybrid technique to safe Internet browsing.

REFERENCES

[1] Anti-Phishing Working Group, "Phishing Activity Trends Report", 4th Quarter April 2014.

[2] Gaurav, Madhuresh Mishra, Anurag Jain "Anti Phishing Techniques: A Review" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 2,Mar-Apr 2012, pp.350-355

[3] A.Naga Venkata Sunil, Anjali Sardana "A PageRank Based Detection Technique for Phishing Web Sites" 2012 IEEE Symposium on Computers & Informatics.

[4] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones "Phishing Detection: A Literature Survey" IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 15, NO. 4, FOURTH QUARTER 2013

[5] Radha Damodaram, Dr. M. L. Valarmathi "Phishing website detection and optimization using Modified bat algorithm" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1, Jan-Feb 2012, pp. 870-876

[6] Atul M. Tonge , Surbhi R. Chaudhari "Phishing Susceptibility and Anti-Phishing Security Strategies-Literature Review" International Journal of Scientific & Engineering Research, Volume 4, Issue 12, December-2013, ISSN 2229-5518

[7] [7] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen, "Detecting Phishing Web sites: A Heuristic URL Based Approach" The 2013 International Conference on Advanced Technologies for Communications.

[8] Amruta Deshmukh, Sachin Mahabale, Kalyani Ghanwat, Asiya Sayyad "WEB PHISH DETECTION (AN EVOLUTIONARY APPROACH)" International Journal of Research in Engineering and Technology eISSN: 2319-1163, pISSN: 2321-7308, Volume: 03

[9] Antonio San Martino, Xavier Perramon "Phishing Secrets: History, Effects, and Countermeasures" International Journal of Network Security, Vol.11, No.3, PP.163–171, Nov. 2010

[10] Tareq Allan, Justin Zhan "Toward Fraud Detection Methodology" IEEE 2010.

[11] Pawan Prakash, Manish Kumar, Ramana Rao Kompella, Minaxi Gupta "PhishNet: Predictive Blacklisting to detect Phishing Attacks" presented as conference at IEEE INFOCOM, 2010

[12] Radha Damodaram, M. L. Valarmathi “Bacterial Foraging Optimization for fake website Detection” TIJCSA, January 2013

[13] Weiwei Zhuang, Qingshan Jiang, Tengke Xiong “An Intelligent Anti-phishing Strategy Model for Phishing Website Detection” 2012 32nd International Conference on Distributed Computing Systems Workshops