International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 02, Issue – 03, March – 2017, PP – 120-129

Detecting Phishing Websites, a Heuristic Approach

Suman Bhattacharyya[1], Chetan kumar Pal[2], Praveen kumar Pandey[3]
[1]*(Department of Information Technology, Xavier Institute of Engineering/Mumbai University, India)*
[2]*(Department of Information Technology, Xavier Institute of Engineering/Mumbai University, India)*
3(Department of Information Technology, Xavier Institute of Engineering/Mumbai University, India)

Abstract: : Phishing is a website forgery technique with an intention to track and steal the sensitive information of online users. The hacker fools the user with social engineering techniques such as SMS, voice, email, website and malware. Various approaches have been proposed and implemented to detect a variety of phishing attacks such as use of blacklists and whitelists to name a few. In this paper, we propose a desktop application called PhishSaver, which focuses on URL and website content of the phishing webpage. We aim at detecting phishing websites with the help of a desktop application named PhishSaver. PhishSaver uses a combination of blacklist and a number of heuristic features to detect a number of phishing attacks. For blacklist, we have used GOOGLE API SERVICES that is Google safe browsing blacklist as this list is constantly updated and maintained by Google. It is also possible to run PhishSaver as a daemon process that means it is able to detect phishing attacks in real time as a user browses the internet. PhishSaver takes URL as input and outputs the status of URL as phishing or legitimate website. The heuristics used to detect phishing are footer links with null value, zero links in the body of the html, copyright content, title content and website identity. PhishSaver is able to detect zero hour phishing attacks which may have not been blacklisted and is faster than visual based assessment techniques that are used in detecting phishing. We observe that PhishSaver has obtained a higher accuracy rateand covers a wider range of phishing attacks that results in less false negative and false positive rate.

Keywords: Anti-phishing; Copyright; Footer Html; Google Safe Browsing; Phishing; Website Identity.

1. INTRODUCTION

PhishSaver is a desktop application to effectively detect phishing websites and practices. PhishSaver is capable to detect most of the phishing techniques and aims at providing full-proof security from all kinds of phishing attacks.PhishSaver is not a traditional antivirus software and does not guarantee any protection from any kind of virus attacks. PhishSaver is based on the idea that users should be able to browse the internet safely and access websites without getting concerned about the legitimacy of the websites. The system works by taking an input from the user in the form of URL of the website for which legitimacy needs to be determined. The system then outputs the state of the website as phishing, legitimate or unknown.

1.1. Why PhishSaver?

- The e-banking phishing websites can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate.
- This application can be used by many E-commerce enterprises in order to make the whole transaction process secure.
- The algorithm used in this system provides better performance as compared to other traditional classification algorithms.
- By using this system user can purchase products online securely.

1.2. Feasibility Study

People often purchase products online and make payment through e-banking. There are many E-banking phishing websites. In order to detect the e-banking phishing website our system uses an effective heuristic algorithm. The e-banking phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria.

- Economic Feasibility
  - This system can be used by the E-commerce enterprise in order to make the whole transaction process securely. This system will increase the productivity and profitability of the E-commerce enterprise. This will provide economic benefits. It includes quantification and identification of all the benefits expected.
- Operational Feasibility

- This system is more reliable, maintainable, affordable and producible. These are the parameters which are considered during design and development of this project. During design and development phase of this project there was appropriate and timely application of engineering and management efforts to meet the previously mentioned parameters.

- Technical Feasibility
- The system uses a safe browsing blacklist database that is a database consisting of registry of URLs of the websites that are verified as phishing or malware hosting websites and their description.
- There is basic requirement of hardware to run this application. This system is developed in Java. This application requires internet connection to perform a lookup online.

1.3. Motivation

The existing systems have several limitations which can be overcome by PhishSaver. The main advantages of PhishSaver over normal phishing detectors are as following:

- It is based on heuristic approach that is capable of detecting Zero hour phishing attacks that is phishing attacks that are relatively new which is not possible for most of the other phishing detectors.

- The system involves just five modules that act as filters to determine the legitimacy of the URL.

- Users just need to provide the URL of the website whose legitimacy needs to be determined. Nothing else needs to be done by the user.

- Even complex phishing attacks can be easily determined by this algorithm. This algorithm has relatively less false positive and false negative rates.

- The expected accuracy rate for PhishSaver for detecting phishing websites is calculated to be 96.57%.

- The main advantage of our application is that it can detect phishing sites which tricks the users by replacing content with images, which most of the existing anti phishing techniques are not able to detect, even if they can, they take more execution time than our application.

2. RELATED WORK

Several different solutions for phishing have been developed during the past few years. These solutions include governmental policies against online frauds, creating awareness to users and technology countermeasures. Review of these researches improved our basic understanding towards this problem and helped us to build a more compressive model for the current study.

2.1 Governmental policies against online frauds

The progressive increase in phishing attacks every year causes financial loss and reputational loss to the individuals and companies. To mitigate this growing number of attacks, number of laws and regulations has been introduced by governing bodies around the globe. These laws have established standards for protecting against illegal use of personal information and increased penalties for criminals involved in phishing. The laws are not within the technical scope of the paper so no details of such laws would be elaborated in this paper. In further section, we look into some technical work carried out in the relevant field.

2.2 Technology based countermeasures to automatically detect phishing attack

In this section, we will review some of the researches offering technological solutions against phishing attack. These researches are broadly grouped into six categories based on the techniques used to assess genuineness of the webpage.

2.3.1 Blacklist based Approaches

Blacklist approaches, are approaches in which a list of known phishing sites is maintained and the website under scrutiny is checked against such list. This blacklist is usually gathered from multiple data sources like spam traps or spam filters, user posts, and verified phish compiled by third parties such as takedown vendors or financial institutions.

2.3.2 White-list based Approaches

In contrast to blacklist approach, white-list approaches maintain list of all safe websites and their associated information. Any website that does not appear in the list is treated as a suspicious website. Most of the white-list approaches are universal white-list which requires all legitimate sites in the world to be included in the list. But it is not easy to maintain all the legitimate sites in the web under one roof to decide the legitimacy of a webpage. There are very few researches that focus on improving the white-list.

2.3.3 Heuristics based Phishing Detection

Heuristic based methods extract features of a webpage to decide the legitimacy of the website instead of depending on any precompiled lists. Most of these features are extracted from URL and HTML Document Object Model (DOM) of the given webpage. The extracted features are compared with known features collected from phishing and legitimate pages to decide its legitimacy. Some of these approaches use heuristics to calculate spoof score of a given webpage to check its genuineness.

2.3.4 Multifaceted Approaches in Identifying Phishing Pages

Multifaceted approaches play a vital role in detection of phishing webpage and phished target it mimics. These techniques use any combination of techniques in computational science to detect phishing websites.

2.3.5 Visual and Image Similarity based Phishing detection

Phishing websites impersonate to look identical to the legitimate websites of financial institutions or other popular companies. These fake websites are constructed with same title, favicon, styles, layouts, images, flash objects, and content of the legitimate pages. Such phished pages differ very minimal from the legitimate page which makes the user incapable to distinguishing from the genuine webpage. A limitation of these approaches is that they require a method to retrieve a website's content robustly. Any distortion in retrieving the content of the webpage leads to false positive.

2.3.6 Empirical Studies on Anti-Phishing Strategies

A category of research focuses on experimental studies to comprehend the significance of implementing anti-phishing strategies. These studies are based on past events of phishing attacks and involves a careful study of such techniques to apprehend such attacks in the future.

Any of the anti-phishing strategies discussed in this paper does not provide complete protection against phishing attacks. To address this problem, new techniques need to be developed in order to detect phishing websites and source of the phishing websites. In this thesis we have focused on developing technology based anti-phishing methodologies that detect phishing webpages automatically. It would not only overcome the weaknesses of other anti-phishing strategies but would also efficiently detect phishing websites and targeted webpage it mimics.

3. Proposed System

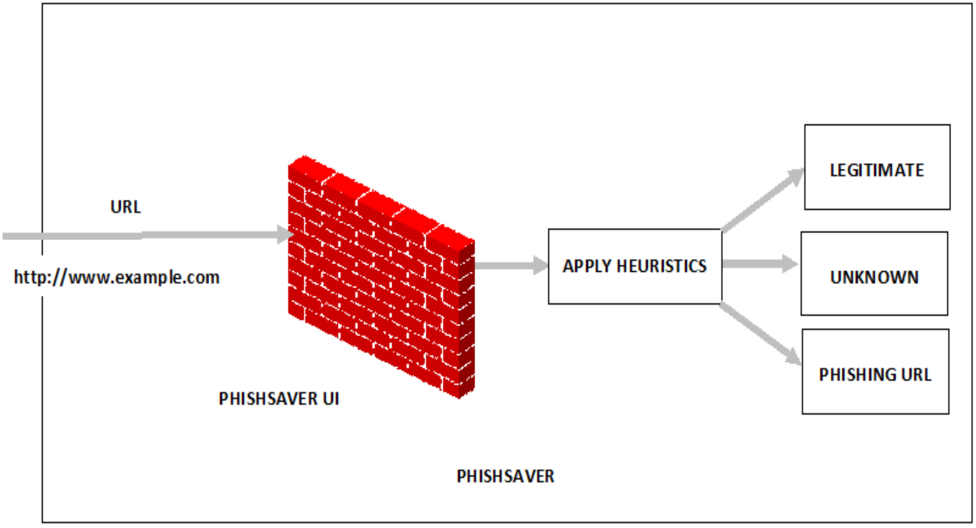


Fig. 1. Architecture of the proposed solution

Figure-1 shows the architecture of our proposed work. We use a combination of blacklist and a number of heuristic features to verify the legitimacy of the URL. For blacklist, we intend to use GOOGLE API SERVICES that is Google safe browsing blacklist as this list is constantly updated and maintained by Google and the heuristic features consists of five modules. These five modules are considered as five levels of detection. PhishSaver takes URL as input and gives output as status of website i.e. phishing or legitimate or unknown.
We also calculated the identity of the URL based on the maximum frequency of domain that are extracted from the hyperlinks of HTML.

The use of blacklist and the five modules of the system is discussed as follows:

3.1. Use of blacklist

In the first level of detection, domain of the URL is compared with a list of known phishing websites to check its legitimacy against the blacklist. For this we have used GOOGLE SAFE BROWSING blacklist as it is reliable and constantly updated list of blacklisted websites. We have used the GOOGLE SAFE BROWSING API version 4 for this. There are two ways for comparison with the list. Either we download cached form of the list for comparing locally or we perform a lookup online. For our system, we perform a lookup online, so we require an internet connection for this lookup. If comparison is successful and a match is found, the website is designated as a phishing site and the algorithm stops at this point. Otherwise the algorithm proceeds with the next module. Before moving ahead, the webpage is parsed and stored as a document object model (DOM) element.

3.2. Detection of login page

Phishers use phishing tool kits in creating fake login forms to steal sensitive information. It is known that online users reveal sensitive information mostly in a login page. Thus, the key to detecting phishing websites is to search for websites with a login page. In the absence of a login page a website cannot be claimed to be a phishing website since there is no way a user could reveal sensitive information. The login page existence is found through parsing the html of website for input type =”password”. In the presence of the password type field, the application PhishSaver continues the execution otherwise stops the execution process as the user does not have a way to enter his/her confidential information. This filtration would prevent phishing detection process on ordinary websites not containing login forms thus reducing the margin of error in the detection process.

3.3. Footer links pointing to NULL (#)

The footer links that have a null value or a null character and that do not link to any other website are called as null footer links. An anchor tag pointing to NULL value is called as NULL anchor. It is an indication of the link redirecting to its own page. From this fact of information we derived third level of detection heuristic. In this third level of detection, we consider footer links of the websitesand specifically the ones that have null values. Phishers mainly wantusers to stay on the login form. So they design the form to consist of such null footer links, leading to users directing continuously to a page consisting login form. Hence, some researchers have considered proportionality of the null links with total number of links for filtering the phishing sites.But there is a twist here as some of the legitimate websites also may have null links such company logo’s that are pointed to null. But it is quite a fact that none of the legitimate sites have footers that have links pointing to null. Hence, from this observation we derived a heuristic factor to filter the phishing sites i.e. if the anchor tag in the footer section is pointing to null i.e.
<a href = “#”>
<a href = “#skip”
<a href = “#content”>
then the URL is treated as Phishing URL otherwise PhishSaver forwards to next level of detection.

3.4. Use of copyright and title content

In the fourth level of detection, the <div> tag containing copyright section and <title> tag containing title content is extracted from DOM object and checked for domain information. It is a general practice for legitimate sites, to include domain information in copyright and title section and so we use this information to detect phishing. The copyright and title content is extracted and tokenized into terms. Each tokenized term is compared with the URL. If there is a match it indicates the copyright and domain information is identical indicating a legitimate website. If there areno matches it means copyright contains the targeted domain information and is different than the URL domain indicating something fishy and the URL is classified as phishing and the parsed content is forwarded to the next filter. The algorithm moves to the next module irrespective of the results of this module i.e. the next module executes compulsorily irrespective of the results.

3.5. Website identity

Website identity is determined based on the frequency of hyperlinks with in the website. In legitimate website, frequency of the hyperlinks pointing to its own domain is high when compared to frequency of the hyperlinks pointing to foreign domain. As phishers try to imitate the behaviour of legitimate sites, they insert the links in their websites pointing to the target domain. This information is used to identify the website identity of the given URL by calculating domain of the link with maximum frequency. If the domain of input URL of PhishSaver application does not match with domain having maximum frequency (website identity) then input URL is considered as phishing site targeting to domain with maximum frequency.

This filter is used not only to detect phishing websites but also identifies phishers target domain that is being imitated. The parsed HTML content is passed through this filter mandatorily even if the phishing has been detected from the above filters so that target website is revealed to the user.

4. EXPERIMENTATION AND RESULTS

"PhishSaver" is based on URL content and Web content of the URL such as null footer links, copyrights and title content and the safe browsing blacklist's API. To Develop the tool, NetBeans 8.2 IDE, Java Compiler, JSoup API and Chrome Driver was used.

The API was used for parsing the html contents of web page and extracting html contents such as the links in footer, copyright, title, CSS etc. Chrome Driver is a third-party driver tool to initiate chrome browser from within the java application.

For blacklist we used Safe Browsing API provided by google to perform a lookup online which acts as the primary means of detection. For phishing URLs for experimentation purpose, we referred to a website named PhishTank. PhishTank is an anti-phishing website where anyone can submit, verify, track or share phishing data. It maintains a phishing archive consisting of valid, unknown, online or offline phishing sites. To evaluate the performance of the PhishSaver application in detecting phishing websites, a total of 250 valid, invalid, offline, online phishing sites URL was gathered from this site.

4.1. Evaluation metrics

In order to calculate the accuracy of their proposed system they used following evaluation parameters.

False Positive (F_{Pos}):

This measures the rate of legitimate sites (L) wrongly classified as phishing sites (P).

$$F_{Pos} = \frac{L \rightarrow P}{(L \rightarrow P) + (L \rightarrow L)}$$

False Negative (F_{Neg}):

This measures the rate of phishing sites (P) wrongly classified as legitimate sites (L).

$$F_{Pos} = \frac{P \rightarrow L}{(P \rightarrow L) + (P \rightarrow P)}$$

True Positive (TPos):

This measures the rate of phishing sites (P) correctly classified as Phishing sites (P).

$$TPos = \frac{P \rightarrow P}{(P \rightarrow P) + (P \rightarrow L)}$$

True Negative (T_{Neg}):

This measures the rate of legitimate sites (L) correctly classified as legitimate sites (L).

$$TNeg = \frac{L \rightarrow L}{(L \rightarrow L) + (L \rightarrow P)}$$

Accuracy (Acc):

This measures the overall rate of correctly detected phishing and legitimate instances in relation to all instances.

$$Acc = \frac{(L \rightarrow L) + (P \rightarrow P)}{(L \rightarrow L) + (L \rightarrow P) + (P \rightarrow L) + (P \rightarrow P)}$$

where L → P is number of legitimate sites misclassified as phishing, L → L is number of legitimate sites correctly classified as legitimate, P → L is number of phishing sites misclassified as legitimate, P → P is number of phishing sites correctly classified as phishing. On experimenting phishing sites and legitimate sites they could get the below values of metrics as shown in Column chart.

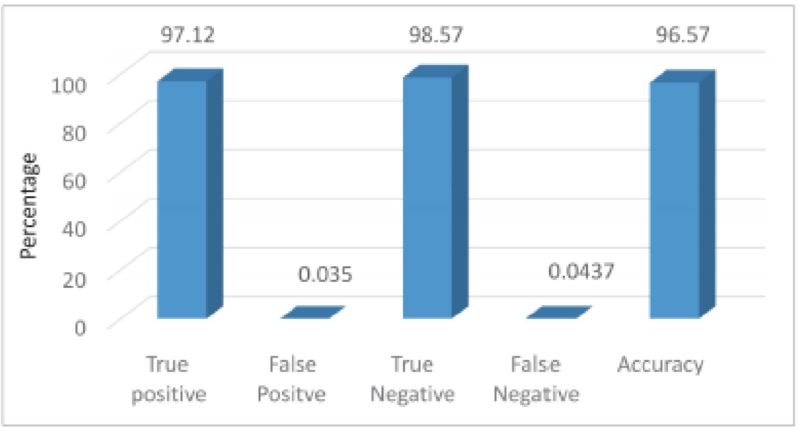
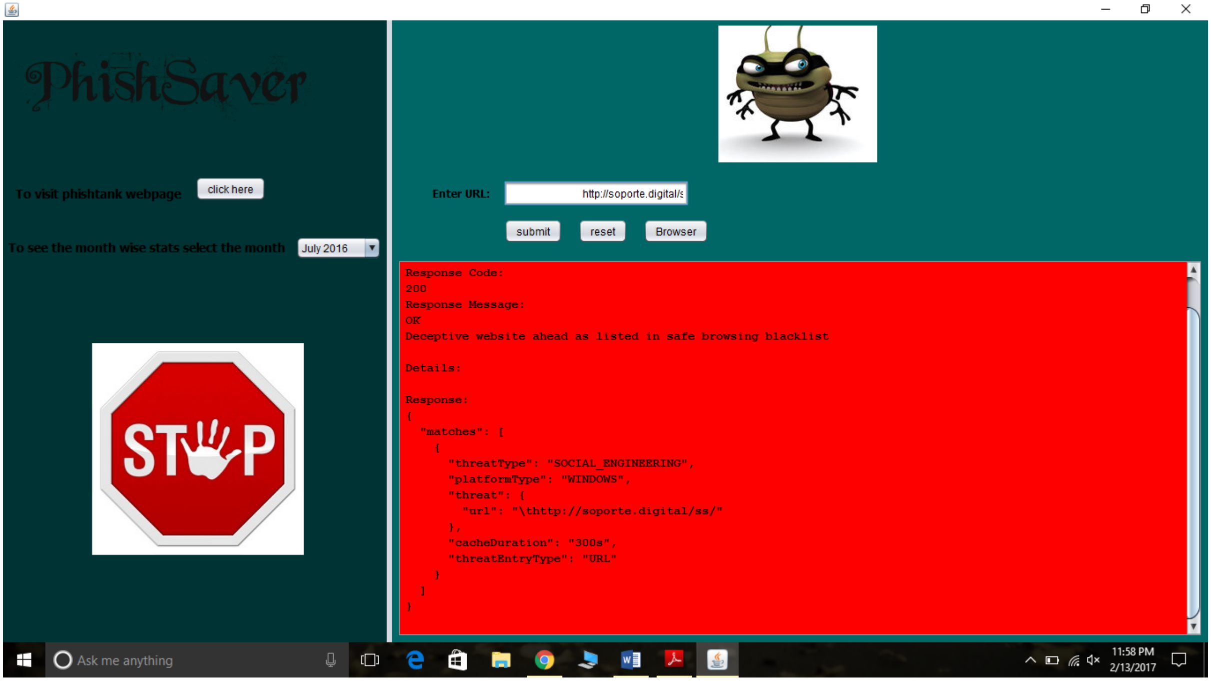


Fig.2. Performance results of PhishSaver application

4.2. Some screenshots of the developed system:

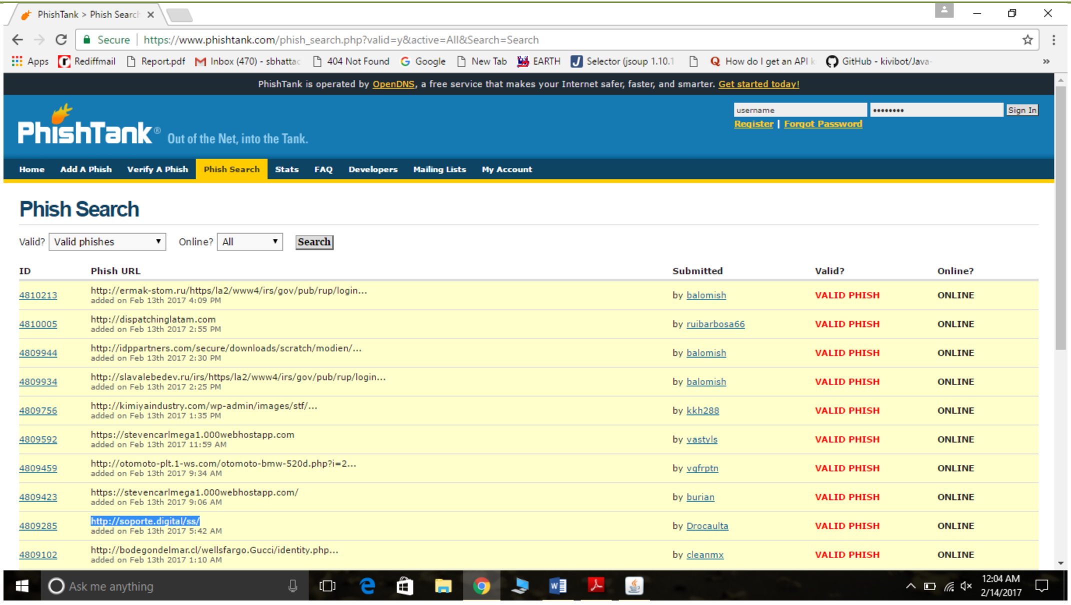


```
Response Code:
200
Response Message:
OK
Deceptive website ahead as listed in safe browsing blacklist

Details:

Response:
{
  "matches": [
    {
      "threatType": "SOCIAL_ENGINEERING",
      "platformType": "WINDOWS",
      "threat": {
        "url": "\thttp://soporte.digital/ss/"
      },
      "cacheDuration": "300s",
      "threatEntryType": "URL"
    }
  ]
}
```

Fig.3 A snapshot of the system detecting a phishing URL

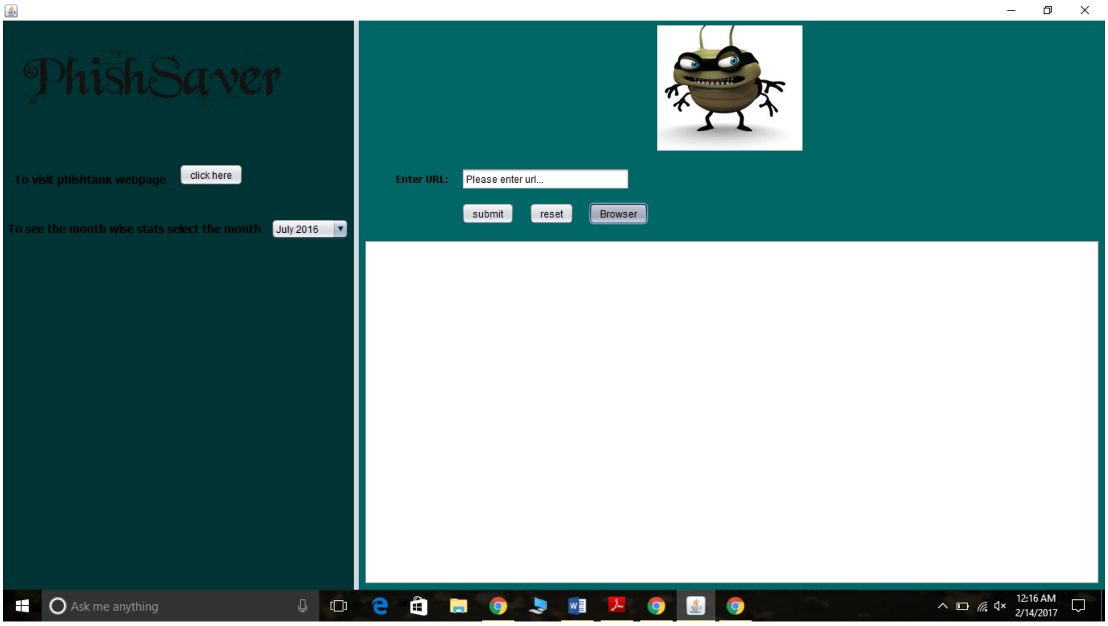

PhishTank > Phish Search
Secure | https://www.phishtank.com/phish_search.php?valid=y&active=All&Search=Search
PhishTank is operated by OpenDNS, a free service that makes your Internet safer, faster, and smarter. Get started today!
PhishTank® Out of the Net, into the Tank.
Register | Forgot Password
Home
Add A Phish
Verify A Phish
Phish Search
Stats
FAQ
Developers
Mailing Lists
My Account

Phish Search

Valid? Valid phishes ▾ Online? All ▾ Search

ID	Phish URL	Submitted	Valid?	Online?
4810213	http://ermak-stom.ru/https/la2/www4/irs/gov/pub/rup/login... added on Feb 13th 2017 4:09 PM	by balomish	VALID PHISH	ONLINE
4810005	http://dispatchinglatam.com added on Feb 13th 2017 2:55 PM	by ruibarbosa66	VALID PHISH	ONLINE
4809944	http://idppartners.com/secure/downloads/scratch/modien/... added on Feb 13th 2017 2:30 PM	by balomish	VALID PHISH	ONLINE
4809934	http://slavalebedev.ru/irs/https/la2/www4/irs/gov/pub/rup/login... added on Feb 13th 2017 2:25 PM	by balomish	VALID PHISH	ONLINE
4809756	http://kimiyaindustry.com/wp-admin/images/stf/... added on Feb 13th 2017 1:35 PM	by kkh288	VALID PHISH	ONLINE
4809592	https://stevencarlmega1.000webhostapp.com added on Feb 13th 2017 11:59 AM	by vastyls	VALID PHISH	ONLINE
4809459	http://otomoto-plt.1-ws.com/otomoto-bmw-520d.php?i=2... added on Feb 13th 2017 9:34 AM	by vqfrptn	VALID PHISH	ONLINE
4809423	https://stevencarlmega1.000webhostapp.com/ added on Feb 13th 2017 9:06 AM	by burian	VALID PHISH	ONLINE
4809285	http://soporte.digital/ss/ added on Feb 13th 2017 5:42 AM	by Drocaulta	VALID PHISH	ONLINE
4809102	http://bodegondelmar.cl/wellsfargo.Gucci/identity.php... added on Feb 13th 2017 1:10 AM	by cleanmx	VALID PHISH	ONLINE

Fig.4 A snapshot of the website Phishtank from which the sample URLs are taken.

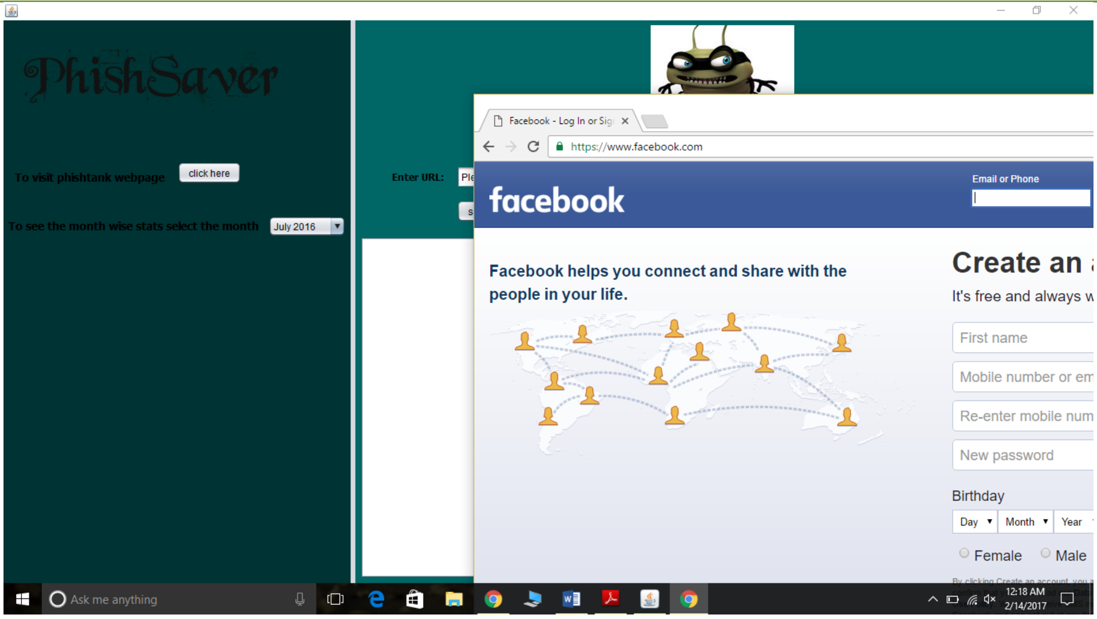


To visit phishtank webpage

To see the month wise stats select the month July 2016

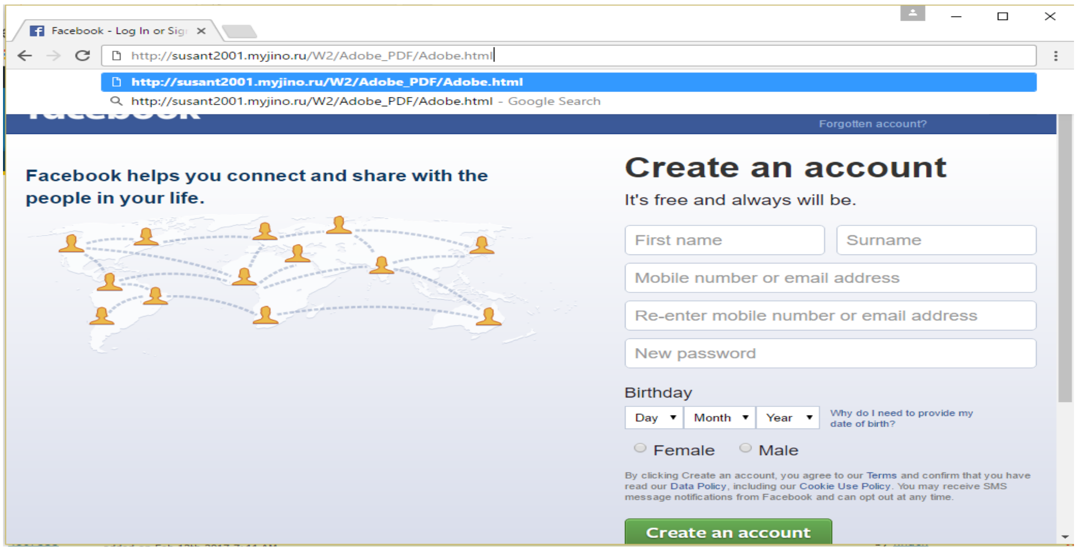
Enter URL: Please enter url...

Fig.5 Clicking the browser button creates a new secure browsing session as shown



Facebook helps you connect and share with the people in your life.

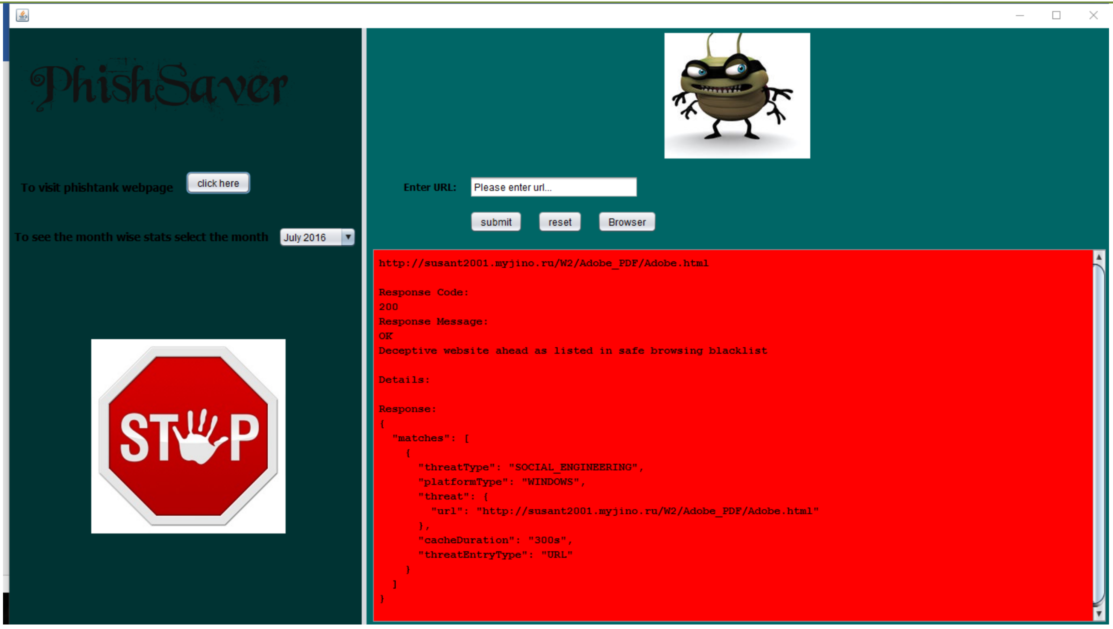
Fig 6. A new safe browser session


Facebook - Log In or Sig
http://susant2001.myjino.ru/W2/Adobe_PDF/Adobe.html
http://susant2001.myjino.ru/W2/Adobe_PDF/Adobe.html
http://susant2001.myjino.ru/W2/Adobe_PDF/Adobe.html - Google Search
Forgotten account?
Create an account
It's free and always will be.
First name
Surname
Mobile number or email address
Re-enter mobile number or email address
New password
Birthday
Day
Month
Year
Why do I need to provide my date of birth?
Female
Male
Create an account

Facebook helps you connect and share with the people in your life.

By clicking Create an account, you agree to our Terms and confirm that you have read our Data Policy, including our Cookie Use Policy. You may receive SMS message notifications from Facebook and can opt out at any time.

Fig 7. Entering a phishing URL results into termination of the browsing session



```
http://susant2001.myjino.ru/W2/Adobe_PDF/Adobe.html

Response Code:
200
Response Message:
OK
Deceptive website ahead as listed in safe browsing blacklist

Details:

Response:
{
  "matches": [
    {
      "threatType": "SOCIAL_ENGINEERING",
      "platformType": "WINDOWS",
      "threat": {
        "url": "http://susant2001.myjino.ru/W2/Adobe_PDF/Adobe.html"
      },
      "cacheDuration": "300s",
      "threatEntryType": "URL"
    }
  ]
}
```

Fig 8. Browser terminated by the application thus preventing from a phishing attack

5. CONCLUSION

A conclusion section must be included and should indicate clearly the advantages, limitations, and possible applications of the paper. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. (10)

Acknowledgements

We are grateful to our project guide Prof.JaychandUpadhyay and we acknowledge his guidance and assistance. Without his encouragement and assistance this project would not have been successful and he helped great role in the development of this project.

REFERENCES

[1]. APWG, Phishing activity trends paper.[online]. http://docs.apwg.org/reports/APWG_Global_Phishing_Report_1H_2014.pdf

[2]. APWG, Phishing activity trends paper.[online]. http://docs.apwg.org/reports/APWG_GlobalPhishingSurvey_2H2013.pdf

[3]. Sophos, Do-it-yourself phishingkits found on the internet, reveals Sophos, Technical paper, Sophos, August (2004). [Online]. http://www.sophos.com/pressoffice/news/articles/2004/08/sa− diyphishing.html

[4]. Safe Browsing API – Google Developer, [Online] Available at https://developers.google.com/safe-browsing/

[5]. P. Prakash, M. Kumar, R. R. Kompella and M. Gupta, Phishnet: Predictive Blacklisting to Detect Phishing Attacks, In INFOCOM, 2010 Proceedings IEEE, pp. 1–5, March (2010).

[6]. Y. Cao, W. Han and Y. Le, Anti-Phishing based on Automated Individual White-List, In Proceedings of the 4th ACM Workshop on Digital Identity Management ACM, pp. 51–60, October (2008).

[7]. Y. Joshi, S. Saklikar, D. Das and S. Saha, PhishGuard: A Browser Plug-In for Protection from Phishing, In 2nd International Conference on Internet Multimedia Services Architecture and Applications, IMSAA 2008, IEEE, pp. 1–6, December (2008).

[8]. Y. Zhang, J. I. Hong and L. F. Cranor, Cantina: A Content-Based Approach to Detecting Phishing Web Sites, In Proceedings of the 16th International Conference on World Wide Web, ACM, pp. 639–648, May (2007).

[9]. N. Chou, R. Ledesma, Y. Teraguchi and J. C. Mitchell, Client-Side Defense Against Web-Based Identity Theft, In NDSS, February (2004).

[10]. A. Y. Fu, L. Wenyin and X. Deng, Detecting Phishing Web Pages with Visual Similarity Assessment based on Earth Mover's Distance (EMD), IEEE Transactions on Dependable and Secure Computing, vol. 3(4), pp. 301–311, (2006)

[11]. J. Mao, P. Li, K. Li, T. Wei and Z. Liang, Baitalarm: Detecting Phishing Sites using Similarity in Fundamental Visual Features, In 5th International Conference on Intelligent Networking and Collaborative Systems, INCoS 2013, IEEE, pp. 790–795, September (2013).

[12]. L. Wenyin, G. Huang, L. Xiaoyue, Z. Min and X. Deng, Detection of Phishing Webpages based on Visual Similarity, In Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, ACM, pp. 1060–1061, May (2005).

[13]. L. Wenyin, G. Huang, L. Xiaoyue, Z. Min and X. Deng, Detection of Phishing Webpages based on Visual Similarity, In 14th International Conference on World Wide Web (WWW): Special Interest Tracks and Posters, (2005).

[14]. G. Xiang, and J. I. Hong, A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval, In Proceedings of the 18th International Conference on World Wide Web, ACM, pp. 571–580, April (2009).

[15]. G. Ramesh, I. Krishnamurthi and K. S. S. Kumar, An Efficacious Method for Detecting Phishing Webpages through Target Domain Identification, Decision Support Systems, vol. 61, pp. 12–22, (2014). [Online]. Available at: http://www.sciencedirect.com/science/article/pii/S0167923614000037