Journal of Engineering and Applied Sciences 12 (Special Issue 7): 8088-8094, 2017
ISSN: 1816-949X

PhishSys: A Honey Bee Inspired Intelligent System for Phishing Websites Detection

Abdulghani Ali Ahmed and Ali Safa Sadiq
University Malaysia Pahang, Pahang, Malaysia

Abstract: Phishing website lures computer users to interact with the spoofed websites rather than the real ones. The main purpose of this attack is to steal the confidential information from the users. The attacker creates a fake webpage that looks similar to the real page. This deceitful act allows the hacker to observe and modify user's confidential information. This study proposes a real time system inspired from honey-bee defence mechanism in nature for filtering phishing website attacks (PhishSys). In particular, the proposed system PhishSys filters phishing website through three main phases of investigation: PhishTank-Match (PM), Undesirable-Absent (UA) and Desirable-Present (DP) investigation phases. The PM technique is used in the first phase in order to check if the requested URL is listed in the blacklist of PhishTank database. The UA technique is used in the next phase of investigation for checking the absence of undesirable symbols in Uniform Resources Locators (URLs) of the requested website. The DP technique is used in the last phase of investigation in order to check the presence of the requested URL in the desirable whitelist. The obtained results show that the detection mechanism is deployable and capable to detect various types of phishing attacks with maintaining a low rate of false alarms.

Key words: Phishing website, URL, honey bee, intelligent system, Desirable-Present (DP), Undesirable-Absent (UA)

INTRODUCTION

Phishing website is a common social engineering attack used to reveal confidential and private information through deceiving the users without being discovered (Ludl *et al*., 2007). The main aim of this attack is to illegally obtain confidential information such as username, password and accounts numbers. In general, phishing website is one instance of social engineering attack. Phishing attack can be launched using several techniques such as SMS, chatting, VOIP and fraudster emails. Users typically have several user accounts on different websites including bank account, e-Mail and also social network accounts. Hence, the most vulnerable targets towards this attack are the innocent web users who are unaware of their valuable information.

A statistical report generated by Anti-Phishing Working Group Organization stated that 163,333 phishing attacks have occurred in 2014. Another report generated by Intel Security (2015) stated that number of new suspicious URLs reached 30,000,000 at the third quarter of 2014. These reports also stated that web phishing was categorized among the top 26% of network threats. Some crime groups practice phishing website attack as a business. Thus, billions of dollars have been reported as stolen amounts from banks in US, Russia and Eastern Europe.

In particular, web spoofing attack uses the social engineering to trap the victim through a fake URL for redirecting him to a phishing website. The spoofed link is sent via. e-Mail to the victim or placed on the popular websites. The phishing webpage is created similar to the original webpage. Thus, the victim request will be directed to the attacker server rather than directing it to the real web server. Figure 1 illustrates the steps involved in phishing website attack.

Many researches are recently conducted to detect phishing website attacks. Nevertheless, these researches are not sufficiently capable to prevent the sophisticated hack of phishing websites. Moreover, the usage of various media communication such as social network is a reason to increase numbers of web attacks. According to Jagatic *et al*. (2007), 70% of phishing website attacks are executed through social network. In fact, the lack of awareness and education on phishing website attack helps the attackers to launch their hack successfully.

Failure to differentiate between the spoofed and real websites is still a challenge in the existing systems of phishing website detection. The existing systems of malware, anti-phishing software do not effectively detect the phishing website attacks. Moreover, the digital Certificate (CA) and Secure Socket Layer (SSL) are

Corresponding Author: Abdulghani Ali Ahmed, University Malaysia Pahang, Pahang, Malaysia

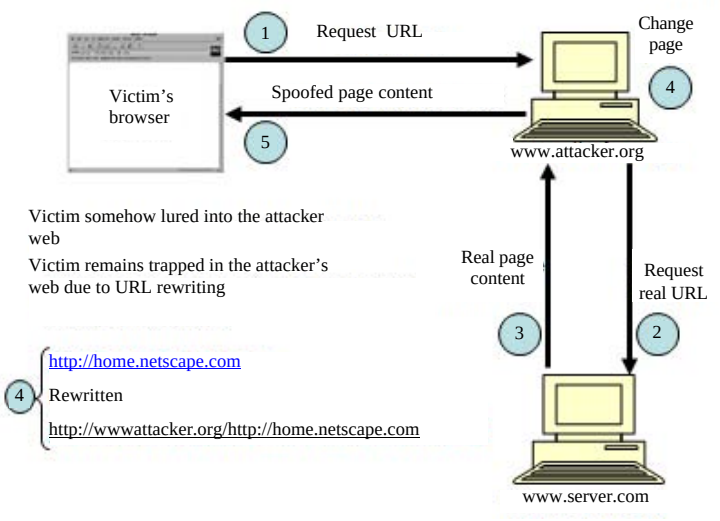

Fig. 1: Steps involved in phishing website attack

unable to immune web users against the real time phishing attacks. In phishing website attack, the attacker diverts the request to fake web server. In fact, certain type of SSL and CA can be forged while everything appears to be legitimate.

In the recent years, many researches such as Rains *et al*. (2008) and Srinoy (2007) have demonstrated that natural insect's behavior system may provide us with a powerful strategies that can be applied to information security. The social system of honeybees on organizing and protecting their colony is one of the feasible strategies that can be inspired to design an effective protection system against network intrusions. In the nature, honeybees survive in risky environments with different levels of threats to security. These threats motivate the bees to obtained practice defense skills to detect and early respond on any action that may threaten the colony (Couvillon *et al*., 2008).

Honeybee defenders face the same challenge as the one faced by phishing website detection system. As phishing detection system faces challenge of differentiation between normal and spoofed websites, honeybee defenders face challenge of differentiation between behaviors of the intruders and the legitimate nest-mate. In the bee's colony, there is a small entrance protected by particular guards. The responsibility of the entrance guards is to examine incomers at the colony entrance and prevent them to enter the colony if they are intruders (Butler and Free, 1951). According to Stabentheiner *et al*., (2002), honeybee guards separate between nest-mates and non nestmates by using two main methods: Undesirable-Absent (UA) and Desirable-Present (DP). Further details about the UA and DP methods are provided by Jantan and Ahmed (2014a, b).

This study develops a real time system for phishing website detection (PhishSys) based on honey-bee defence mechanism in nature. In particular, the proposed system PhishSys filters phishing website through three main phases of investigation: PhishTank-Match (PM), Undesirable-Absent (UA) and Desirable-Present (DP). The PM technique is used in the first phase in order to investigate the URL of the suspicious website and filtered it as a phishing website if it is included in PhishTank blacklist (PhishTank, 2015). In the UA phase, URL of the requested website is further investigated and filtered as a phishing website if any undesirable symbol presents in its content. The DP technique is used in the third phase to investigate if the suspicious URL which is not existed in PhishTank blacklist and does not include any undesirable feature is not a new phishing website. In the DP phase, the suspicious URLs will be inspected based on the whitelist of websites URLs. Figure 2 shows the architecture of the proposed system PhishSys of phishing websites detection.

Literature review: This study discusses the most related research of phishing website attacks. Several researches on phishing website attack have been conducted during the past few years. Many studies and researches have been conducted to detect phishing website attacks. The researches of phishing website attack prevention are survived and classified into various approaches: content-based, heuristic-based and blacklist-based approaches as shown in Fig. 3.

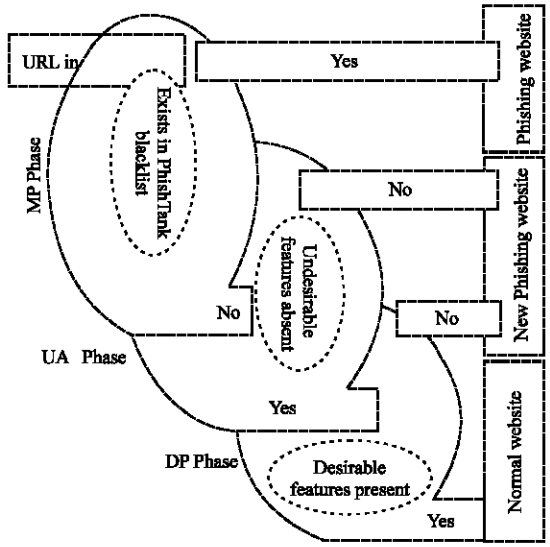

Fig. 2: PhishSys architecture

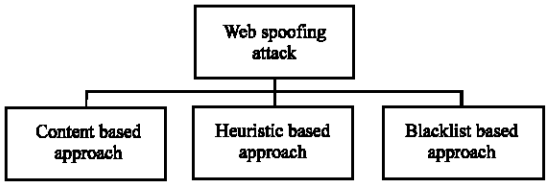

Fig. 3: Web spoofing attack detection approaches

MATERIALS AND METHODS

Content based approach: The similarity based index among original website and the spoofed ones could be identified using content approaches. This type of approaches calculates the similarity between two web pages based on the matches of the web page content. Generally speaking, we can achieve adequate level of accuracy with low false alarms (in identifying the fake web page) by using such type of spoofing detection approaches. The researcher (Zhang *et al*., 2007) have proposed CANTINA approach as its working based on content similarity in identifying phishing web sites. The phishing websites were detected by using Term Frequency/Inverse Document Frequency (TF-IDF). Utilizing CANTINA approach, the false positive rate was successfully reduced. That was due to the use of TF-IDF technique to retrieve information and text mining. The detection level of CANTINA approach has shown that it's able to catch about 97% phishing websites with about 6% false positive alarm. The researchers (Zhang *et al*., 2007) also have reported that with merging heuristics approach to TF-IDF, it could manage to catch around 89% of phishing sites with only 1% false positives.

Although, CANTINA approach could effectively recognize the phishing website, it has some limitations in dealing with text hidden in HTML. Attacks of hidden text in HTML may evade the keyword extraction technique due to such limitation. Moreover, CANTINA approach inquires for wider-scale deployment and evaluation to improve its functionality in phishing website detection.

On the other hand, the researchers (Dunlop *et al*., 2010) have proposed a GoldPhish approach which is considered under a content-based category. This approach utilizes Google as a search engine in identifying fishing websites. The main assumption of this approach is that fake websites are typically active for short period of time. GoldPhish approach works on taking images for the active websites in the user's web browser. Afterwards, optical character recognition technique was utilized to convert the collected images into computer text form. As a way to identifying the possibility of phishing attacks and analyzing the page rank, GoldPhish approach will then make use of the transformed text as an input into the search engine. Based on the performance analysis of the proposed GoldPhish approach we could observe that it was effective in mitigating the false positive and it could detect new phishing website. In spite of this improvement, GoldPhish approach has its limitation in terms of time delay in exploring webpages. Besides, this approach could be vulnerable to attacks on Google's PageRank algorithm as well as Google's search service.

Heuristic based approach: In this study, we discuss the heuristic based approach that elaborates HTML or URL signature in identifying the spoofed webpages. Many researchers have conducted studies on phishing website detection based on this type of approaches. For instance, by Chou researcher have introduced a SpoofGurad as one of the techniques that uses heuristics approach. SpoofGurad technique is considered as an anti-phishing browser plug-ins. A combination of stateless page evaluation has been implemented in this technique for state full page evaluation. Moreover, this technique computes the spoof index value as a way to examine the outgoing post data. The researchers have identified a threshold value that helps in classifying the webpages into phishing/normal pages when the calculated spoof index is more than this value the user will be notified about this page. On the other hand when the spoof index value is less than the pre-defined threshold value, the webpage will be classified as legitimate page. Although, SpoofGuard technique has a drawbacks as it produces high rate of false positive alarms in case new phishing attack.

On the other hand (Ludl *et al*., 2007; Sheng *et al*., 2009), the webpage structure phishing URLs are analyzed

to discriminate among legitimate and phishing web pages. These studies are relying on distinguish the features of the individualities that could be used in perceiving the phishing web pages. Utilizing these types of website phishing detection techniques, phishing attack could be recognized and conveyed as soon as it is activated. Thus, it will assist in degrading the need to sustain a blacklist that could inquire time in identifying and produces more complexity. Nevertheless, these techniques are also produces high false negative rate due to the fact that there are numerus phishing websites which are categorized as legitimate.

Blacklist based approach: As another strategy that is widely used now a days in identifying phishing websites, blacklist based approach has been implemented as anti-phishing technique. Using this approach the system should obtain an up to date blacklist for the common phishing websites. Altogether entries that are denied access are allocated in such phishing blacklist (Sheng *et al*., 2009; Ahmed and Abdullah, 2016). Accordingly, users are avoided to access websites that exist in the blacklist. The core role in blacklist based approach is tracing the URLs from phishing websites as a way to preserve and generate the blacklist. These URLs could be traced out from the users phishing emails, spam, or from the associations that aid the anti-phishing such as Anti-Phishing Working Group (APWG) and Phish Tank (PhiskTank, 2015). As soon as a URL is stated will then verify it before it is inserted into the blacklist.

As a representative of phishing detection techniques based on blacklist approach, Net craft Toolbar (Cranor *et al*., 2007) is discussed to provide better understanding. This toolbar discovers the website's security threat using certain criteria such as time of sitting the Net craft web server survey, timestamps of accessing the website, country that hosted the website, name of organization that hosting the current site besides it rates the risk scale.

Using Net craft toolbar approach the possibilities of phishing website attack could be minimized. Moreover, this toolbar could also assist in preventing the users from auto download malicious files that would be used by the phishers in gathering user's private information. Likewise, Net craft can defend the user from the DNS poisoning and shield the user from the pop-up windows that could hide the address bar. In spite of the positive points behind this approach in defending the user, the user might face new types of phishing attack. The blacklist database involves a endless updating in order to aware the new discovered URLs of the phishing websites.

Proposed model PhisFilter: Phishing website attacks happen when the user is directed to the spoofed website using fake URLs. This study describes the agents and algorithm of PhishSys as a proposed system for phishing attack detection. In particular PhishSys consists of three main agents: PM agent, UA agent and DP agent. Further details about PhishSys agents are provided in the following subsections.

PM agent: This agent is used to investigate the suspicious URL by checking if it exists in the blacklist of PhishTank database. The PhishTank is one of the public databases of phishing URLs (https://www.phishtank.com/). The PhishTank API accepts an HTTP POST request along with the query URL and returns a Java Script Object Notation (JSON) object in response which tells whether the query URL is phishing or not. The use of JavaScript code is to check the suspicious URL links either it is secure page or suspicious. The function of checking URL is to send a POST request to PhishTank in order to use their API. The function will then process the returned response and determine whether the page is a verified phishing website or not by looking at the response's result. The function would then return true if the URL is verified as belonging to PhishTank blacklist or false if not.

UA agent: This agent is responsible to further investigate the content of the URL which was filtered by PM agent in the previous phase. Its main aim is to check if the URL contains one of the undesirable features. According to Osareh and Shadgar (2008), undesirable features can be used to differentiate between legitimate and phishing web pages.

These undesirable features are checked using several criteria such as IP Address, long URL address, adding a prefix or suffix, redirecting using the symbol "//" and URLs having the symbol "@". These features are inspected using a set of rules in order to distinguish URLs of phishing webpages from the URLs of legitimate websites. Below is a description for some of these rules.

Some URLs of phishing web page have an addition at the front of the real URLs. An example of this addition is http://www.legitimate.com//http://www.phishing.com. This feature checks the location of the symbol "//" in the URL. If the URL starts with "HTTP", this means that symbol "//" should appear in the sixth position. However, if the URL employs "HTTPS" then the symbol "//" should appear in the seventh position.

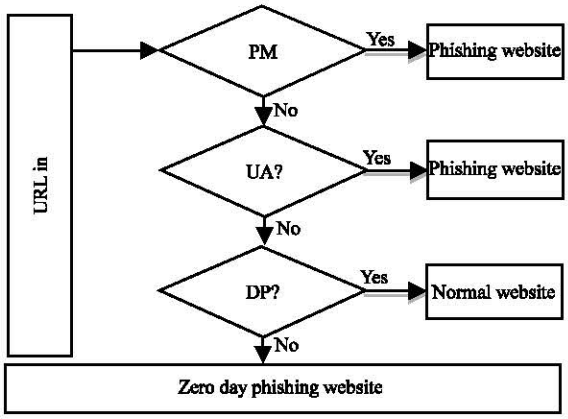

Fig. 4: PhishSys algorithm

DP agent: Given the features of the zero-day attack usually not included in the signature databases, a potent defense system should use alternative strategy to detect the new attacks. PhishSys system provides an agent for further investigating the suspicious URL which does not exist in the blacklist of PhishTank database and verifying if that URL is a new phishing website. This is done by checking the whitelist of legitimate URLs which is the opposite of a blacklist and contains a list of known trusted sites. More detail about the anti-phishing methods using whitelist database is described by Garera *et al.* (2007).

PhishSys uses the mentioned agents in order to achieve an integrated process for filtering the phishing websites. Figure 4 illustrated the main methodology of PhishSys system.

RESULTS AND DISCUSSION

Implementation and results: PhishSys system is implemented and tested in Google Chrome Explorer. Chrome extension allows adding functionality in Google Chrome without diving deeply into native code. Now a days, Chrome provides extension with many purpose APIs for developer to enhance the user browsing experience. This method require Chrome platform APIs to perform its function. Those platforms applied in this method are JavaScript APIs Manifest file format and permission warnings. JSON-formatted manifest file, named manifest.json was included as part of the Chrome extension package.

The use of JavaScript code is to check the URL link either it is secure page or suspicious URL site. The function of checkURL is to send a POST request to PhishTank in order to use their API. The function will then process the returned response and determine whether the page is a verified phishing website or not by looking at the response's result.verified attribute. The function would then return true if the URL is verified as belonging to a phishing website and false if not. The PhishTank database is one of the familiar public crowd sourced database of phishing URLs contain blacklist data collection. The PhishTank API accepts an HTTP POST request along with the query URL and returns a JSON object in response which tells whether the query URL is phishing or not.

Table 1: Examples of phishing websites

Normal sites	Fake websites/Phishing websites
Facebook.com	http://appssecure.at.au/facebook.html
Google drive	http://scredlble.com/login/GDrive/1ff71c3539 f0ba9ab 99eeb75fe36c5a2/
Paypal.com	http://support-paypai.com.itunesverificationhelp. ga/signin/webapps/6282b/websrc
Adobe.com	http://onlinksoft.org/dragon1/products/adobe.php? email=abuse@the-fat-slags.co.uk
Alibaba.com	http://www.footballnewsheadlines.co.uk/wp-admin/css/ alibaba/index.php?email=abuse@gmai.com
Bankofamerica.com	http://integral.rs/log/verify.html
Netflix.com	http://ebayproduct.com/fonts/UpdateService/netbi/ netflixo/Login/payment.php
Outlook.com	http://access0000.wapka.mobi/index.xhtml
Google doc	www.lighthousebd.info/BG/index.php
Dropbox.com	http://dropx.allon4dallas.com/71cbb335c02b4e4c 65c7cb74bef95278/
Amazon.com	http://triofloridashow.com/ap/amzon/amzon/2baf777d 7c13d97bbeb7f3fbcdafb07c/index/web/login.php? action=billing_login=true

The outcome of this method is displaying pop up alert as the result to notify user either the targeted website is secure page or suspicious URL site. When the page is safe, Chrome will notify "safe" page only in the console and redirect to the targeted page allowing user to continue browsing. If the URL website is suspicious, Chrome will notify the user with pop up alert with message "This is a suspected phishing page!!!" This to make user aware the targeted page is a suspicious website that probably a phishing site that luring user to give their personal information.

For detecting URLs phishing website by using chrome extension, the extension is loaded into the user's browser as shown in Fig. 5.

We used blacklist method to detect phishing website by using PhishTank database. Here, we tested 11 suspected phishing sites using this method. Table 1 shows a list of valid phishing URL used in this experiment.

Results testing and evaluation: This study tests proposed system PhishSys is to verify its ability of identifying the phishing website. To this end, a list of 100 URLs is used (59 legitimate web pages and 41 fake web pages). The chosen URLs are randomly selected from the

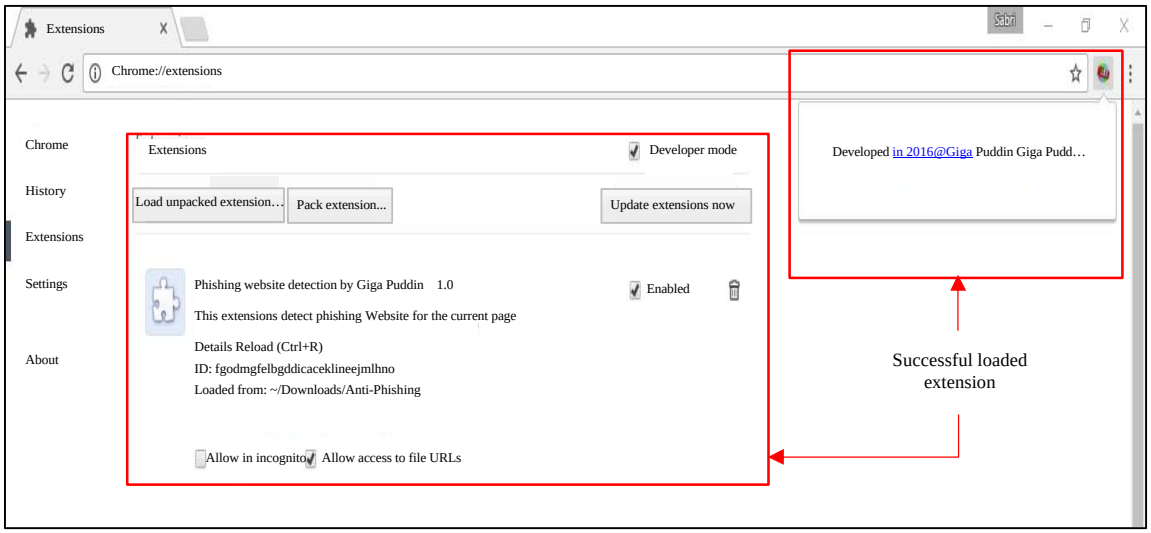

Fig. 5: PhishSys extension in google chrome

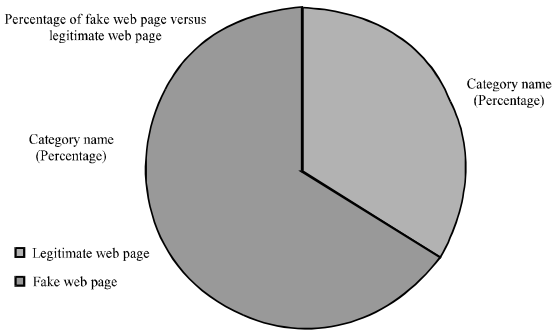

Fig. 6: Phishing detection accuracy

PhishTank (2015). For every URL, PhishSys checks whether the URL has the features of the phishing web page or not.

The obtained result demonstrates that PhishSys classifies 68 of the URLs as legitimate websites and classifies the other 32 URLs as phishing websites. Figure 6 illustrates that 68% from the tested URLs are filtered as legitimate web pages and 32% are filtered as phishing websites.

CONCLUSION

Phishing website attack is difficult to detect. Even with the newest protection systems, these attacks still happen. This study proposes a system to differentiate between the legitimate and phishing web pages. Finding of this research demonstrates its ability to identify the fake webpages based on their URLs. The advantage of PhishSys is its ability to mitigate the limitations of the existing works. PhishSys mitigate the high false rate by using several phased of investigation. On the other hand, PhishSys avoid the high overhead against system performance by considering the filtration strategy. In PhishSys, no need to investigate all requested URLs, instead, only the suspicious ones will be filtered for investigation. Furthermore, investigation the suspicious URL against the whitelist websites is only required for a short listed URLs to detect the zero-day phishing websites.

RECOMMENDATIONS

Future research of this study will improve the method of investigating the URL content which is the UA agent. Investigation the undesirable features will be improved using an intelligent method which will train the system to detect the undesirable features based on the desirable and undesirable patterns.

ACKNOWLEDGEMENTS

RDU grant number RDU160106, Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang supported this research.

REFERENCES

Ahmed, A.A. and N.A. Abdullah, 2016. Real time detection of phishing websites. Proceedings of the IEEE 7th Annual Conference on Information Technology Electronics and Mobile Communication (IEMCON) 2016, October 13-15, 2016, IEEE, Pahang, Malaysia, ISBN: 978-1-5090-0997-8, pp: 1-6.

Butler, C.G. and J.B. Free, 1951. The behaviour of worker honeybees at the hive entrance. Behav., 4: 262-291.

Couvillon, M.J., E.J. Robinson, B. Atkinson, L. Child and K.R. Dent *et al*., 2008. En garde: Rapid shifts in honeybee, Apis mellifera, guarding behaviour are triggered by onslaught of conspecific intruders. Anim. Behav., 76: 1653-1658.

Cranor, L.F., S. Egelman, J.I. Hong and Y. Zhang, 2007. Phinding Phish: An Evaluation of Anti-Phishing Toolbars. Carnegie Mellon University, Pittsburgh, Pennsylvania,.

Dunlop, M., S. Groat and D. Shelly, 2010. Goldphish: Using images for content-based phishing analysis. Proceedings of the IEEE 5th International Conference on Internet Monitoring and Protection, May 9-15, 2010, Barcelona, pp: 123-128.

Garera, S., N. Provos, M. Chew and A.D. Rubin, 2007. A framework for detection and measurement of phishing attacks. Proceedings of the 5th ACM Workshop on Recurring Malcode, October 29-November 2, 2007, Alexandria, VA., USA., pp: 1-8.

Intel Security, 2015. Millions of mobile app users are still exposed to SSL vulnerabilities. Intel Security, Santa Clara, California, USA. https://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q4-2014.pdf.

Jagatic, T.N., N.A. Johnson, M. Jakobsson and F. Menczer, 2007. Social phishing. Commun. ACM, 50: 94-100.

Jantan, A. and A.A. Ahmed, 2014b. Honey bee intelligent model for network zero day attack detection. Intl. J. Digital Content Technol. Appl., 8: 45-45/.

Jantan, A. and A.A. Ahmed, 2014a. Honeybee protection system for detecting and preventing network attacks. J. Theor. Appl. Inf. Technol., 64: 38-47.

Ludl, C., S. McAllister, E. Kirda and C. Kruegel, 2007. On the effectiveness of techniques to detect phishing sites. Proceedings of the 4th International Conference on Detection of Intrusions and Malware and Vulnerability Assessment, July 12-13, 2007, Lucerne, Switzerland, pp: 20-39.

Osareh, A. and B. Shadgar, 2008. Intrusion detection in computer networks based on machine learning algorithms. Intl. J. Comput. Sci. Netw. Secur., 8: 15-23.

PhishTank, 2015. Join the fight against phishing. PhishTank, San Francisco, California, USA. https://www.phishtank.com/.

Rains, G.C., J.K. Tomberlin and D. Kulasiri, 2008. Using insect sniffing devices for detection. Trends Biotechnol., 26: 288-294.

Sheng, S., B. Wardman, G. Warner, L.F. Cranor and J. Hong *et al*., 2009. An empirical analysis of phishing blacklists. Proceedings of 6th Conference on Email and Anti-Spam (CEAS) 2009, July 16-17, 2009,: Carnegie Mellon University, Mountain View, California,-pp: 1.

Srinoy, S., 2007. Intrusion detection model based on particle swarm optimization and support vector machine. Proceedings of the IEEE Symposium on Computational Intelligence in Security and Defense Applications, April 1-5, 2007, Honolulu, HI., pp: 186-192.

Stabentheiner, A., H. Kovac and S. Schmaranzer, 2002. Honeybee nestmate recognition: The thermal behaviour of guards and their examinees. J. Exp. Biol., 205: 2637-2642.

Zhang, Y., J. Hong and L. Cranor, 2007. CANTINA: A content-based approach to detecting phishing web sites. Proceedings of the 16th International Conference on World Wide Web, May 8-12, 2007, Banff, Alberta, Canada, pp: 639-648.