Survey on Malicious URL Hitches, Propagation Mechanisms and Analysis of Classification Algorithms

Samridhi Sharma[#1], Shabnam Parveen[*2]

[#]*M.tech, Dept. of Computer Science and Engineering*
Seth Jai ParkashMukandLal Institute of Engineering and Technology
KurukshetraUniversity, India
[*]*Assistant Professor, Dept. of Computer Science and Engineering*
Seth Jai ParkashMukandLal Institute of Engineering and Technology
Kurukshetra,University, India

***Abstract*-Malicious URL detection has become increasingly difficult due to the evolution of phishing campaigns and efforts to avoidweakening blacklists. The existing state of cybercrime has allowed pirates to host campaigns with smaller lifespan, which reduces the efficacy of the backlist. At the same time,standard supervised learning algorithms are known to generalize in specific patterns observed in the training data, which makes them a better alternative against piracy campaigns. However the highly dynamic environment og these campaigns requires models updated frequently, which poses new challenge as most learning algorithms are too computationally require exclusive retraining. This paper surveys two contributions. Firstly it discusses the problems associated with Malicious URL and there propagation mechanism. Secondly, it provides method to detect and distinguish Malicious URL by analyzing them.For analysis Recall, Precision and F-measures matrices are used.**

***IndexTerms*—Attacks, Adware Classification, Malicious web page analysis, Malicious URLs, Machine Learning.**

I. INTRODUCTION

Ad ware, short for Malicious Software advertising [1] is a sequence of instructions that perform malicious activities on a computernet. The antiquity of malware initiated with the term "computer virus", a term introduced by Cohen. This is a piece of code that replicates by attaching itself to other executable in the system. Today, the malware includes viruses, worms, Trojans, root kits, backdoors, bots, spyware, adware, scare ware and any other program that has malicious behavior. Adware is a fast risingdanger to current computer networks. Manufacturing of Adware hasnow become a multi-billion. The development of the Internet, the arrival of social networks and the rapid proliferation of botnets has caused an exponential increase in the extent of Adware. In 2010, there was a drastic upsurge in the amount of Adware spread through spam emails sent machines that were part of botnets. McAfee Labs reported that there were 6 million new infections each month. [2]

A web malware mentions to each malware that uses the internet to enable cybercrime. In exercise, web malwares could use several kinds of malware and fraud. A public feature is that web malwares all use HTTP or HTTPS protocols, nevertheless a little malwares could additionally use supplementary protocols and constituents, such as links in emails or IMs, or malware attachments. Across web malwares, cyber-criminals regularly rob trustworthy data or hijack computers as bots in botnets. It has been well comprehended that web malwares lead to huge dangers, encompassing profitablecharges, individuality thefts, overthrows of trustworthy data, thefts of web resources, broken brand and confidential standing, and erosion of customer assurance in e-commerce and online banking. Although the exact adversarymechanisms behind web convict hobbies could vary, they all endeavor to bait users to sojourn malicious websites by clicking a corresponding URL (Uniform Resource Locator). A URL is shouted malicious (also recognized as black) if it is crafted in a malicious intention and leads a user to a specific malware that could come to be an attack, such as spyware, malware, and phishing. Malicious URLs are a special choice on the web. Therefore, noticing malicious URLs is a vital task in web protection intelligence.

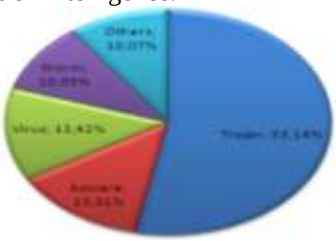

Fig1: Distribution of Types of Malwares

Trojans once again represent the category of malware that has grown most, accounting for 53.14% of the whole. Interestingly, traditional viruses also appear to be making a comeback in recent months and have risen 10 points over the last two quarters, now considering for 24.35% of all new malware [3].In exercise, malicious URL detection faces countless challenges.

A. *Realtime detection*:

To protect users efficiently, a user ought to be cautioned beforehand she/he visits a malicious URL. The malicious URL detection period ought to be extremely short so that users should not have to pause for long and tolerate from poor user experience.

B. *Detection of new URLs:*

To circumvent being noticed, attackers frequently craft new malicious URLs frequently. Therefore, a competent malicious URL detection method has to be able to notice new, unseen malicious URLs. In exercise, the skill of noticing new, unseen malicious URLs is of particular significance, as rising malicious URLs regularly have higher hit counts, and could cause bigcompensations to users.

C. *Competent detection:*

The detection ought to have an elevated accuracy. The precision is of concern; the halt frequency of URLs ought to additionally be considered. From a user's point of think, the accuracy of a detection method is the number of periods that the detection method classifies a URL accurately versus the number of periods that the method is consulted. Gratify note that a URL could be dispatched to a detection method several periods, and ought to be counted several periods in the accuracy calculation. Therefore, noticing oftentimes visited URLs accurately is important. Similarly, it is exceedingly desirable that a malicious URL detection method ought to have an elevated recall so that countless malicious URLs can be detected. Again, after recall is computed in this context, the sojourn frequency of URLs ought to be considered.

To encounter the above trials, the latest malicious URL detection methods endeavor to craft a classifier established on URLs. Anessentialhypothesis is that a spotless training example of malicious URL and good URL examples is available. Such methods segment a URL into tokens employing a little delimiters, such as "/" and "?", and use such tokens as features. A little method additionally removesadditional features, such asWHOIS data and geographic properties of URLs. Then, contraption discovering methods are requested to train an association ideal from the URL sample.

II. MALICIOUS URLPROPAGATION MECHANISMS

The early memorable URL shortening ability was TinyURL that was dispatched in 2002. Its accomplishment enticed competitors and nowadays, there are hundreds of disparate URL shortening services that sporadically proposal supplementary features, as a method of differentiating themselves from the rest. When a user visits a URL, there browser is automatically redirected to the destination page, generally across the use of applicable HTTP rank memos (HTTP 301 or 302), or supplementary client-side mechanisms, e.g., JavaScript or HTML Meta tags. By the alike period, the URL shortening ability lists the sojourn and creates aggregate statistics concerning the visitors that clicked on every single exact short URL, that are usually made available openly or just to the creator of the short link.

A. Ad-based URL shortening services:

Ad-based URL shortening [4] services are services that use advertising and referral plans to enthuse users to craft and allocate short links by paying them a tiny number of money for every single sojourn to their short URLs. For the user who generates the short link, the procedure is comparable to shortening a link alongside each supplementary URL shortening service. The key difference is that the link-creating users have to report together with the ability, if user wants to become salaried for the traffic that user afterward brings.

B. Static page and redirection:

Whenever one more user clicks on the link shortened by an ad-based URL shortening ability, user fields on the service's "Waiting Page", whereas user have to early discern an advertisement for at least an insufficient seconds beforehand user is allowed to continue to the final destination of the short URL. The top portion of the page is manipulated by the ad-based URL shortening ability and the bottom one presents the promoted content inside an iframe. The timed "Continue" button becomes alert and clickable merely afterward a predetermined number of seconds. This ensures that the link-following user gets exposed to the ad beforehand tolerating to the landing page. Across this period span, the landing page's URL is not revealed. Reliant on the ability, it might be plainly obfuscated, or loaded asynchronously from the service's server by a JavaScript routine. A little service additionally use the top portion of the page to display supplementary publicizing banners, maximizing the screen real-estate dedicated to ads.

C. Advertised page:

The iframe displaying the ad to the user is below the maximum manipulation of the advertiser. Barring the use of present HTML5 tags that check the functionality obtainable to the page inside an iframe, an advertiser is free to run arbitrary JavaScript program, Flash, and Java requests, set cookies on the visitor's browser, and display arbitrary content. Finally, note that the ads occurring after a user follows a short URL are random, and depend on every single package's inner presenting arrangement as well as the available ads. Thus, there is no assurance that after two users pursue the alike short URL, that they will be exposed to alike advertisement.

D. iFrame Redirections:

As mentioned earlier, ad-based URL services place advertisements in a frame that spans most of the "Waiting Page" that the user encounters when clicking on a short link. The usage of an frame adequatelysplits the advertiser from the including page, since the advertising scripts cannot access the DOM of the parent frame due to the Same-Origin Policy (SOP) [5], ainfluential security mechanism imposed by all browsers. The SOP, however, does not stop the attacker from redirecting the entire page to an arbitrary destination. This can be easily done in JavaScript by simply setting the toplocation variable to the desired destination URL. This technique is called "frame-busting" and has been associated with sites that tried to protect themselves against click jacking, an attack built on version a prey page in an invisible iframe overlaying a malign page, and

attracting the user to interact with the malicious page. Legitimate sites would include (and still do) a simple JavaScript snippet which would detect the fact that they were “framed" and escape the iframe, as follows:

In ad-based URL services, though, it is the untrusted party that is trapped and can present the precise alike check, escaping the iframe and redirecting the whole tab of the user's browser. Thus, an attacker can redirect the victim from the service's “Waiting Page", to browser-exploiting pages, scams and phishing attacks. Interestingly, attackers can use their maximum manipulation to conduct extra urbane phishing attacks. For instance, as, by default, a locale rendered in an iframe has maximum admission to JavaScript and plugins, the attacker can fingerprint the user's browser and redirect merely specific users to a phishing location, i.e., conduct a spear-phishing attack. Moreover, for the locations that disclosure the page's short URL to advertisers an attacker can notice to that locate the user will be redirected after user clicks the shortening service's time activated button, and can therefore redirect the destination site.

Finally, because of the period that the user needs to pause beforehand she is allowed to continue to the landing page, fluctuating from 5 to 10 seconds for the learned services, it is probable that the user will switch focus to one more tab, therefore not observing the redirection to a phishing page. As conflicted for in the tabnabbing attack this defeat of focus can raise the chances that the user will afterward trust that the phishing page is a legitimate one, and continue to reveal her credentials. Even present browsers contain iframe-restricting mechanisms that permit a parent page to harshly restrict the manipulation of an attacker, inappropriately, none of the examined services are presently employing them discern.

III. RELATED WORK

Malicious URL has become an important Internet security concern. Attackers has attracted towards the online social media due to the openness and accessibility of vast data on these media, to conduct phishing attacks, inject malicious codes, spread malware, and unveiling drive-by-download attacks. To identifydifferent types of malware, S.Divya et al. [6] Study the categories of malware, their vulnerabilities and the existing handling mechanisms. Their study concludes two parameters false positive rate and infection ratio in detecting the malware. Yossi Spiegel etal., [7] discover the choice amid vending new multimedia commercially and bundling it alongside ads and allocating it for free as adware. To sold the software commercially only when its perceived quality is high. It display that adware is extra lucrative after the observed quality of the multimedia is moderately low. In [8], by taking the example of PDF, it suggests the use of HTTP request from a PDF can be attractive for an attacker.An attacker can well force the victim toaccess some malicious web pages.HodaEldardiry et al.[9] has proposed a malicious insiders detection prototype which includes two types of activities blend-in anomalywhere malicious insiders try to behave similar to a group they do not belong. For this behavioral inconsistencies across these domains are observed which include logon, device, file, http, email sent and email received, and unusual change anomaly where malicious insiders exhibit changes in theirbehavior. Fusion algorithm is used to combine anomaly from multiple source of information.William T.Young et al., [10] this paper presents the realistic associate menace instances in a real company database of computer custom activity. Area vision is requested (1) to select appropriate features for use by structural anomaly detection algorithms, (2) to recognize features indicative of attention recognized to be associated alongside associate menace, and (3) to ideal recognized or distrusted instances of associate menace scenarios.Neha Gupta et al., [11]have implemented the concept of URL shortners. A shortening service Bitly is used with the dataset of 763,160 short URLs.Their study concludes that it is not using spam detection services efficiently. For detecting malicious URL,URL and two domain specific features are collected and conclude the comparative results by achieving 86.41%accuracy. Luca Invernizziet.al, [12] present EVILSEED approach to search for the web pages that are malicious and concluded that this approach is efficient than crawler based approaches. Jian Cao et al., [13] have focused on forwarding based features along with URL and graph based features in order to train a detection model. They assess the arrangement employing concerning 100,000 early memos amassed from SinaWeibo, which is the biggest OSN website in China. Their study concludes that the forwarding base features are more effective than conventional features because of the high accuracy and low false positive rate.Da Huang et al., [14] they have stated two points. Firstly they counsel to vibrantly remove lexical outlines from URLs. Second, they develop a new process to source their novel URL outlines that are not assembled employing each pre-defined items their comprehensive empirical discover employing the real data sets from Fortinet, a head in the web.Nick Nikiforakis et al., [15] this paper examines the ecosystem of ad-based URL shortening services. They argue that due to the monetary incentives and the attendance of third-party publicizing webs, ad-based URL shortening services and their users are exposed to extra hazards than established shortening services.BirhanuEshete et al., [16] they tackle the setback of noticing whether a given URL is hosted by an exploit kit. They use machine learning approach to detect the malicious URL. Comprehensive examinations alongside real globe malicious URLs expose that WEBWINNOW is exceedingly competent in the detection of malicious URLs hosted by exploit kits alongside extremely low false-positives.HeshamMekky et al., [17] developed a methodology to recognize malicious shackles of HTTP redirections. Then, they apply a supervised decision tree classifier to recognize malicious chain which results recall and precision benefits above 90% and up to 98%. Karan B. Maniar[18]has shown that there are many different types of cyber security threats, but at the same time, there are numerous

ways to avert those threats. H. B. Kazemian et al., [19] has proposed several machine learning models for text classification to classify the web pages as either malicious or not. There resultsconcluded 89% supervised learning and 87% for unsupervised algorithms.

IV. ANALYSIS OF EXISTING TECHNIQUES FOR MALICIOUS URL DETECTION.

TableI: Exisiting Techniques.

AUTHORS	WORKED ON	TECHNIQUES USED	RESULTS
ValentinHamon [8]	PDF language and security Model	PDF Objects and Java Scripts.	Detected Malicious Code in PDF documents
Da Huang et.al[14]	Pattern mining	Complete Pattern set, Greedy Algorithm.	Rum Time and No of Patterns Increases Graduallyr both the Algorithms
William t.young et.al [10]	Introduced a Language for Specifying Anomaly	Grid Based Fast AnomalyDscovery given Duplicates.	Accuracy-99.5 percentile
BrihanuEshete et.al[16]	Kit Workflows	J48, Random Forest Logistic Regression.	Accuracy-99.7%

In our survey work done we noticed that for every malicious URL detection firstly we have to make a repository of features set following which we can classify the URL as malign or begnin (safe URL). The larger the features set we build more correctly we can classify Malicious URL. Nowadays malicious URL comes in many forms such as short URL (concept of URL shortners), long URL, content based, irrelevant links, images, ad based URL. To address this issue many algorithms have been proposed. These algorithms function differently with different feature sets. Clustering algorithm or classification algorithm is used in some of the papers to detect the malicious URLs and some has used the combination of both. When both are used the results produced is much better. This is so because it diminishes the time taken to group the huge multi-dimensional dataset and categorizes them accurately.

Size of the dataset also plays a crucial role in detecting the malicious URLs. The factors include such as error rate must be lowest. Our survey works include that there are some Performance parameters by which the result for a given feature set is concluded. In this section we have analyzed that malicious URL can be in any form and there are numbers of techniques to detect them on the basis of the features we chose and the technology wise.

V. ANALYSIS OF CLASSIFICATION ALGORITHMS USING DIFFERENT PARAMETERS

To check whether classification algorithms are working accurately or not we choose three performance parameters: Precision, Recall and F-measures. These are the elementaryprocedures and using this performance matrix is formed. Different algorithms attain different level of performance.

In this table shown below we have showed the performance by taking the value between 0 and 1.

- High is for the value equal to 1
- Medium is for value between 0.6 to 0.9
- Low is for value between 0.0 to 0.5

Table II Analysis of Classification Algorithms.

Classification techniques	Precision	Recall	F-measures
Decision Tree	Medium	Low	Low
Neural Networks	High	Low	Medium
Naïve Bayes	Medium	Medium	Medium
Support Vector Machine	High	Low	Low
Random Forest	Medium	Medium	Medium

On the basis of these performance parameters an evaluation has been made that which algorithm will produce best results. These parameters are further categorized into four types which include True positive rate, True negative rate, false positive rate and false negative rate.

Performance Evaluation:

A. *True positive rate*-It is number of real positive instance which are classified correctly as positive.
B. *True negative rate*-It is number of real negative instance which are classified correctly as negative.
C. *False positive rate*- It is number of real negative instance which are classified incorrectly as positive.
D. *False negative rate*- It isnumber of real positive instance which are classified incorrectly as negative.

With these four values precision, recall and f-measures are measured.

1). Precision: It is the number of precisely classified instance of a target class, i.e., positive class, over the number of instance classified as view to that class. It is also known as positive predicted value.

$$Precision = \frac{TPR}{TPR + FPR}$$

Where TPR=True Positive Rate.
FPR=False Positive Rate.

*2). Recall:*It is the number of precisely classified instance of a class, i.e., positive class, over the number of instance of that class. The other name for recall is sensitivity.

$$Recall = \frac{TPR}{TPR + FNR}$$

Where FNR= false Negative Rate.

3).F-measure: The F-measure can be viewed as a compromise between recall and precision. It is high only when both recall and precision are high. It is the harmonic mean of Precision and recall.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

VI. CONCLUSION AND FUTURE SCOPE

This survey has presented a number of malicious URL problems along with mechanism of their propagation.URL classification is an important information retrieval task. Precise classification of search queries benefits a number of higher-level tasks such as Web search and ad matching. Through overall research we conclude that output of detecting Malicious URL differs because of the different feature set used. In the earlierwork donenumerous machine learning algorithms are used for, automatic generation of classification rules by examining a set of training examples Labeled.Two key steps in the classification are to select the Features to Be Examined and the decision rule to classifythesefeatures based on characteristics. Different clustering and classification techniques are present to classify URL as malign or begnin URLs. In future we can use the combination of clustering and classification technique to increase the speed of detection along with the use of advantages of the neural network to classify the malicious URL efficiently.

REFRENCES

[1]. Apte, Jitendra, and Marina Lima Roesler. "*Interactive multimedia advertising and electronic commerce on a hypertext network."* U.S. Patent No. 7,225,142. 29 May 2007.

[2]. Ravula, Ravindar Reddy. *Classification of Malware using Reverse Engineering and Data Mining Techniques.* Diss. University of Akron, 2011.

[3]. *"Pandalabs Q2 Report Details New Tabnabbing Phishing Scam"*, By PandaSecurity,http://www.pandasecurity.com/mediacenter/news/pandalabs-q2-report-details-new-tabnabbing-phishing-scam/, July 1, 2010.

[4]. Nikiforakis, Nick, Federico Maggi, GianlucaStringhini, M. ZubairRafique, WouterJoosen, Christopher Kruegel, Frank Piessens, Giovanni Vigna, and Stefano Zanero. "*Stranger danger: exploring the ecosystem of ad-based URL shorteningservices."* In Proceedings of the 23rd international conference on World wide web, pp. 51-62. International World Wide Web Conferences Steering Committee, 2014.

[5]. Karlof, Chris, Umesh Shankar, J. Doug Tygar, and David Wagner. "*Dynamic pharming attacks and locked same-originpolicies for web browsers.*" In Proceedings of the 14th ACM conference on Computer and communications security, pp. 58-71. ACM, 2007.

[6]. S. Divya, *"A Survey on Various Security Threats and Classification of Malware Attacks, Vulnerabilities and DetectionTechniques."* International Journal of Computer Science & Applications (TIJCSA) 2, no. 04 (2013).

[7]. Yossi Spiegel, "*Commercial software, adware, and consumer privacy."* International Journal of Industrial Organization 31, no. 6 (2013): 702-713.

[8]. ValentinHamon, *"Malicious URI resolving in PDF documents*." Journal of Computer Virology and Hacking Techniques 9, no. 2 (2013): 65-76.

[9]. HodaEldardiry, Evgeniy Bart, Juan Liu, John Hanley, Bob Price, and Oliver Brdiczka. "*Multi-domain informationfusion for insider threat detection."* In Security and Privacy Workshops (SPW), 2013 IEEE, pp. 45-51. IEEE, 2013.

[10]. William T.Young, Henry G. Goldberg, Alex Memory, and James F. Sartain. *"Use of domain knowledge to detect insiderthreats in computer activities."* In Security and Privacy Workshops (SPW), 2013 IEEE, pp. 60-67. IEEE, 2013.

[11]. Neha Gupta, AnupamaAggarwal, and PonnurangamKumaraguru. "bit. ly/malicious: *Deep Dive into Short URL based e-Crime Detection."* In Electronic Crime Research (eCrime), 2014 APWG Symposium on, pp. 14-24. IEEE, 2014.

[12]. Luca Invernizzi et.al "*EVILSEED: A Guided Approach to Finding Malicious Web* Pages",2012 IEEE 2012 IEEE Symposium on Security and Privacy

[13]. Jian Cao, Qiang Li, YuedeJi, Yukun He, and Dong Guo. *"Detection of Forwarding-Based Malicious URLs in Online Social Networks."* International Journal of Parallel Programming (2014): 1-18.

[14]. Da Huang, Kai Xu, and Jian Pei. *"Malicious URL detection by dynamically mining patterns without pre-defined elements."* World Wide Web 17, no. 6 (2014): 1375-1394

[15]. Nick Nikiforakis, Federico Maggi, GianlucaStringhini, M. ZubairRafique, WouterJoosen, Christopher Kruegel, Frank Piessens, Giovanni Vigna, and Stefano Zanero. "*Stranger danger: exploring the ecosystem of ad-based URL shortening services."* In Proceedings of the 23rd international conference on World wide web, pp. 51-62. International World Wide Web Conferences Steering Committee, 2014.

[16]. BirhanuEshete and V. N. Venkatakrishnan."WebWinnow: "*lveraging exploit kit workflows to detect malicious urls."* In Proceedings of the 4th ACM conference on Data and application security and privacy, pp. 305-312. ACM, 2014.

[17]. HeshamMekky, Ruben Torres, Zhi-Li Zhang, SabyasachiSaha, and Antonio Nucci. "*Detecting malicious HTTP redirections using trees of user browsing activity."* In INFOCOM, 2014 Proceedings IEEE, pp. 1159-1167. IEEE, 2014.

[18]. Karan B. Maniar"Overview of Cyber Security" International Journal of Engineering Trends and Technology (IJETT)Volume 15 Number 3 – Sep 2014

[19]. H. B. Kazemian and S. Ahmed. "*Comparisons of machine learning techniques for detecting malicious webpages."* Expert Systems with Applications 42, no. 3 (2015): 1166-117.