Enhanced Framework for Detecting Malicious HTTP Redirections with Supporting Classifier

Thella Vineela¹, Bondili. Balvinder Singh²

¹,²Computer Science Engineering, Vignan’s Lara Institute of Technology & Science, Guntur, India

Abstract: Malicious URL detection has become increasingly difficult due to the evolution of phishing campaigns and efforts to avoid weakening blacklists. The existing state of cybercrime has allowed pirates to host campaigns with smaller lifespan, which reduces the efficacy of the blacklist. 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 of these campaigns requires models updated frequently, which poses new challenges 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.

Keywords: Attacks, Adware Classification, Malicious web page analysis, Malicious URLs, Machine Learning.

1. Introduction

Adware, short for Malicious Software advertising [1] is a sequence of instructions that perform malicious activities on a computer. 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 rising danger to current computer networks. Manufacturing of Adware has now 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, cybercriminals regularly rob trustworthy data or hijack computers as bots in botnets. It has been well comprehended that web malwares lead to huge dangers, encompassing profitable charges, 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 adversary mechanisms 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.

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.

1. 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.

2. 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 big compensations to users.

3. 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.

Volume 4 Issue 12, December 2015

www.ijsr.net


International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Indexed in: Research Gate, Google Scholar, EBSCO, CrossRef

Keywords: Attacks, Adware Classification, Malicious web page analysis, Malicious URLs, Machine Learning.
To encounter the above trials, the latest malicious URL detection methods endeavor to craft a classifier established on URLs. An essential hypothesis 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 removes additional features, such as WHOIS data and geographic properties of URLs. Then, contraption discovering methods are requested to train an association ideal from the URL sample.

2. Malicious URL Propagation Mechanisms

The early memorable URL shortening ability was Tiny URL 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.

1.Ad-based URL shortening services

Ad-based URL shortening services are services that use advertising and referral plans to entice 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.

2.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.

3.Adverted 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.

4.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 iframe adequately splits 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], a influential 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 top location 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 prepay 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 un trusted 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.

3. Related Works

Malicious URL has become an important Internet security concern. Attackers have 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 identify different 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 et al., [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 to access some malicious web pages. Hoda Eldardiry et al.[9] has proposed a malicious insiders detection detection prototype which includes two types of activities blending anomaly where 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 their behavior. 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 shorteners. A shortening service Bitly is used with the dataset of 763,160 short URLs. Their study concludes that it is not. There results concluded 89% supervised learning and 87% for unsupervised algorithms.

4. Analysis of Existing Techniques for Malicious URL Detection.

A simple Question Answering system was implemented using the technique called Semantic Role Labeling. The system consisted of three phases called Query Processing, Document Processing, Answer Processing. In each of these phases, several language processing components were used. The system is web based and hence uses the search engine to extract information from the web.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Worked On</th>
<th>Techniques Used</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valentín Hamo n</td>
<td>PDF language and security Model</td>
<td>PDF Objects and Java Scripts.</td>
<td>Detected Malicious Code in PDF</td>
</tr>
<tr>
<td>Da Huang et al.</td>
<td>Pattern mining</td>
<td>Complete Pattern set, Greedy Algorithm.</td>
<td>Run Time and No of Patterns Increases Gradually both the Algorithms</td>
</tr>
<tr>
<td>William Young et al</td>
<td>Introduced a Language for Specifying Anomaly</td>
<td>Grid Based Fast Anomaly Discovery given Duplicates.</td>
<td>Accuracy- 99.5 percentile</td>
</tr>
<tr>
<td>BrihanuEshete et al</td>
<td>Kit Workflows</td>
<td>J48, Random Forest Logistic Regression</td>
<td>Accuracy- 99.7%</td>
</tr>
</tbody>
</table>

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 begin (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.

5. 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 elementary procedures 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>
<thead>
<tr>
<th>Classification techniques</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
</tbody>
</table>

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 is number 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.

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.

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

\[ \text{Precision} = \frac{TPR}{TPR + FPR} \]

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

\[ \text{Recall} = \frac{TPR}{TPR + FNR} \]

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.

\[ F - \text{measure} = \frac{2 \times \text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \]

6. 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 earlier work done numerous 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 classify these features based on characteristics. Different clustering and classification techniques are present to classify url as malign or begin in url. 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.

References


Author Profile

Thella Vineela
Obtained the B.Tech degree in Information Technology (IT) Department at Vignan’s Lara Institute Of Technology & Science, Guntur in 2013. At present pursuing the M.Tech in Computer Science and Engineering(CSE) Department at Vignan’s Lara Institute Of Technology & Science, Guntur.

Bonil Balvinder Singh
Obtained the B.Tech Degree in Computer Science and Engineering(CSE) from Malineni Lakshmaiah Engineering College, Singarayakonda in 2008 and M.Tech from VR Siddhartha Engineering College in 2013. He has 3 years of teaching experience and Currently working in Computer Science and Engineering(CSE) Department at Vignan’s Lara Institute Of Technology & Science, Guntur.