MALWARE AND MALICIOUS URL IDENTIFICATION BY USING MACHINE LEARNING

Sushma Pedineedi[1], Shaik Masthanbi[2], Vanukuri Vasantha[3], P Mounika[4], Godavarthi Amar Tej[5]

[1,2,3,4] *B. Tech students, Dept. of ECE, VVIT, Nambur, Guntur District, Andhra Pradesh*
[5] Asst .Prof., Dept. Of ECE , VVIT, Nambur , Guntur District, Andhra Pradesh

---***---

Abstract –*Malicious web sites largely promote the growth of internet criminal activities like stealing one's personal information such as, account details, passwords. This leads to strong financial loss. As a result, there has been strong motivation to develop systemic solution to stop the user from visiting such websites. To avoid the user from visiting such websites we try to propose a learning based approach. In this approach websites are classified into 3 classes. They are Benign, Malware and Malicious. We find that phishing website prefers to have longer URL, more levels(delimited by dot). Malware websites could pretend to be a benign one by containing popular brand names as tokens. And malicious sites are always less popular than benign ones. So, we can consider the site popularity as an important feature. Our proposed technique analyses the Uniform Resource Locator (URL) itself without accessing the content of Web Sites. By employing suitable learning algorithm, we try to achieve better performance on generality and coverage compared with black-listing service.*

Key Words: Black-listing service, Benign, Malware and Malicious.

1. INTRODUCTION

The aim of a phishing attack is to fraudulently acquire sensitive information by masquerading as a legitimate entity in an electronic communication. It attempts to trick users to obtain specific information for financial or other gain, such as credit card/financial details, account passwords, or other personal valuable information. They all attempt to lure users to visit malicious websites by clicking a corresponding URL (Uniform Resource Locator).

2. EXISTING SYSTEM

Blacklisting is a popular process used by all of the major web browsers, that typically warn users about potential harm that can be caused by visiting a webpage that is included in their a-priori blacklist listings. However, using preselected lists may not work with previously unseen URLs, since it is non-trivial to predict the malicious nature of a webpage that has not been visited before.

3. METHODOLOGY

In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

1. Address Bar based Features

If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Rule: IF{ If The Domain Part has an IP Address → Phishing Otherwise → Legitimate

2. Using Pop-up Window

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

3. Long URL to Hide the Suspicious Part

Phishes can use long URL to hide the doubtful part in the addressbar.Forexample:http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html

To ensure accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

Rule: IF{$URLlength$ <54 → $feature$=Legitimate $elseif URLlength$ ≥54 and ≤75 → $feature$=$Suspiciousoth erwise$→ $feature$=Phishing

We have been able to update this feature rule by using a method based on frequency and thus improving upon its accuracy.

4. Using URL Shortening Services "TinyURL"

URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an

"HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL. For example, the URL "http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4".

Rule: IF{Tiny URL → Phishing Otherwise→ Legitimate

5. URL's having "@" Symbol

Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

Rule: IF {URL Having @ Symbol→ Phishing Otherwise→ Legitimate.

6. Redirecting using "//"

The existence of "//" within the URL path means that the user will be redirected to another website. An example of such URL is:"http://www.legitimate.com//http://www.phishing.com". We exam in the location where the "//" appears. We find that if the URL starts with "HTTP", that means the "//" should appear in the sixth position. However, if the URL employs "HTTPS" then the "//" should appear in seventh position.

Rule: IF {The Position of the Last Occurrence of "//" in the URL > 7→ Phishing Otherwise→ Legitimate

7. Adding Prefix or Suffix Separated by (-) to the Domain

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example http://www.Confirme-paypal.com/.

Rule: IF {Domain Name Part Includes (−) Symbol → Phishing Otherwise → Legitimate

8. Sub Domain and Multi Sub Domains

Let us assume we have the following link: http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (CCTLD), which in our example is "uk". The "ac" part is shorthand for "academic", the combined "ac.uk" is called a second-level domain (SLD) and "hud" is the actual name of the domain. To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then, we have to remove the (CCTLD) if it exists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as "Suspicious" since it has one sub domain. However, if the dots are greater than two, it is classified as "Phishing" since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign "Legitimate" to the feature.

Rule: IF {Dots in Domain Part=1 → Legitimate Dots In Domain Part=2 → Suspicious Otherwise→ Phishing

9. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough. Checking the certificate assigned with HTTPS including the extent of the trust certificate issuer, and the certificate age. Furthermore, by testing out our datasets, we find that the minimum age of a reputable certificate is two years.

Rule: IF {Use https and Issuer Is Trusted and Age of Certificate≥ 1 Years → Legitimate Using https and Issuer Is Not Trusted → Suspicious Otherwise→ Phishing

10. Domain Registration Length

Based on the fact that a the phishing website lives for a short period of time, we believe that the trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

Rule: IF {Domains Expires on≤ 1 years → Phishing Otherwise→ Legitimate

11. Favicon

A favicon is a graphic image (icon) associated with a specific webpage. Many existing user agents such as graphical browsers and newsreaders show favicon as a visual reminder of the website identity in the address bar. If the favicon is loaded from a domain other than that shown in the address bar, then the webpage is likely to be considered a Phishing attempt.

Rule: IF{Favicon Loaded From External Domain→ Phishing Otherwise→ Legitimate.

12. The Existence of "HTTPS" Token in the Domain Part of the URL

The phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

Rule: IF{Using HTTP Token in Domain Part of The URL→ Phishing Otherwise→ Legitimate

13. Request URL

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.

Rule: IF {% of Request URL <22% → Legitimate %of Request URL≥22% and 61%→ Suspicious Otherwise→ feature=Phishing.

14. URL of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as “Request URL”. However, for this feature we examine:

1. If the <a> tags and the website have different domain names. This is similar to request URL feature.

2. If the anchor does not link to any webpage, e.g.:

A. <a href=“#”>

B. <a href=“#content”>

C. <a href=“#skip”>

D. <a href=“JavaScript ::void(0)”>

Rule: IF{% of URL Of Anchor <31% → *Legitimate*% of URL Of Anchor ≥31% And≤67% → Suspicious Otherwise→ Phishing

15. Links in <Meta>, <Script> and <Link> tags

Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script; and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage.

Rule: IF{% of Links in "<Meta>","<Script>" and "<Link>"<17% → Legitimate% of Links in <Meta>","<Script>" and "<Link>" ≥17% And≤81% → Suspicious Otherwise→ Phishing

16. Server Form Handler (SFH)

SFHs that contain an empty string or “about: blank” are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.

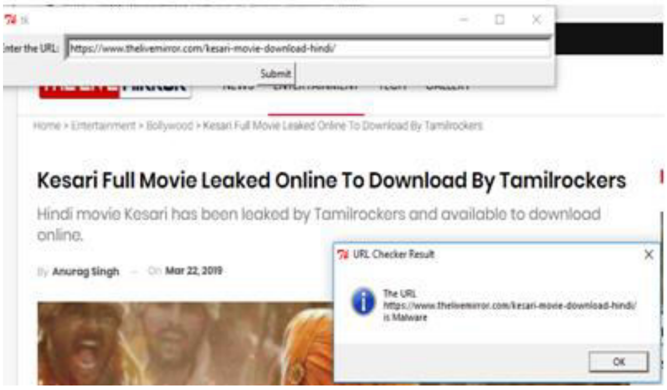
Rule: IF{SFH is "about: blank" Or Is Empty → Phishing SFH Refers To A Different Domain→ Suspicious Otherwise → Legitimate

17. Submitting Information to Email

Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the user’s information to his personal email. To that end, a server-side script language might be used such as “mail()” function in PHP. One more client-side function that might be used for this purpose is the “mailto:” function.

Rule: IF{Using "mail()" or "mailto:" Function to Submit User Information→ Phishing Otherwise → Legitimate.

RESULT



Enter the URL: https://www.theliveemirror.com/kesari-movie-download-hindi/

Home › Entertainment › Bollywood › Kesari Full Movie Leaked Online To Download By Tamilrockers

Kesari Full Movie Leaked Online To Download By Tamilrockers

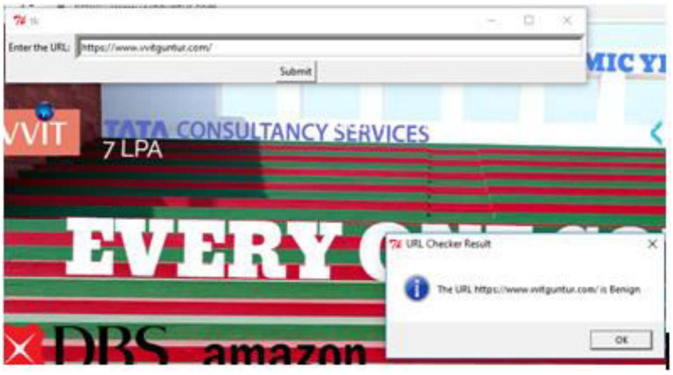
Hindi movie Kesari has been leaked by Tamilrockers and available to download online.

By Anurag Singh — On Mar 22, 2019

URL Checker Result

The URL https://www.theliveemirror.com/kesari-movie-download-hindi/ is Malware

BENGIN

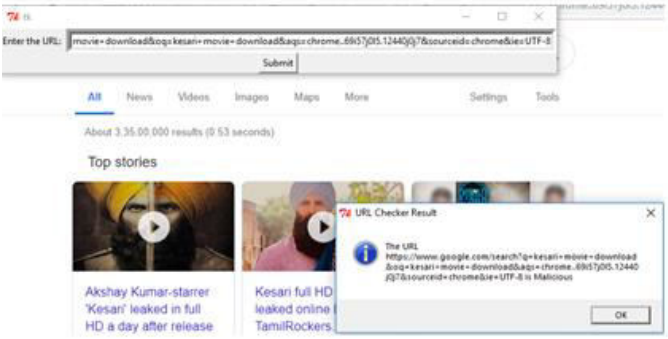


Enter the URL: https://www.vvitguntur.com/

URL Checker Result

The URL https://www.vvitguntur.com/ is Benign

MALICIOUS



Enter the URL: movie+download&oq=kesari+movie+download&aqs=chrome..69i57j0l5.12440j0j7&sourceid=chrome&ie=UTF-8

About 3,35,00,000 results (0.53 seconds)

Top stories

Akshay Kumar-starrer 'Kesari' leaked in full HD a day after release

Kesari full HD leaked online TamilRockers

URL Checker Result

The URL https://www.google.com/search?q=kesari+movie+download&oq=kesari+movie+download&aqs=chrome..69i57j0l5.12440j0j7&sourceid=chrome&ie=UTF-8 is Malicious

CONCLUSION

By using the project “URL Phishing Analysis”, we can reduce the cyber crime activities. It can be recommended to avoid financial and personal data loss. As this project identifies nature of Website without accessing the contents of Website.

ACKNOWLEDGEMENT

The authors can acknowledge any person/authorities in this section.

REFERENCES

[1] Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection and analysis of drive-by-download attacks and malicious java script code", *Proceedings of the 19th International Conference on World Wide Web*, pp. 281-290, 2010.

[2] R. B. Basnet, A. H. Sung, "Mining web to detect phishing urls", *Proceedings of the International Conference on Machine Learning and Applications*,vol. 1, pp. 568-573, Dec 2012.

[3] Mohiuddin Ahmed, AbdunNaserMahmood, Jiankun Hu, "A survey of network anomaly detection techniques", *J. Netw. Comput.Appl.*,vol. 60, no.C, pp. 19-31, 2016.

[4] S. CarolinJeeva, Elijah Blessing Rajsingh, "Intelligent phishing url detection using association rule mining", *Human-centric Computing and Information Sciences*, vol. 6, no. 1, pp. 10, 2016.