Malicious URL Detection using Machine Learning Algorithms

Marcelo Ferreira

Lusofona University of Porto, Portugal
ferreira_marcelo@outlook.pt

Abstract. Malicious URLs are a dangerous threat to cyber security, these types of attacks can lead to scams, where people lose money, their information and accounts. It is important to be able to detect and act against these threats, the most conventional way is the use of blacklists, but this technique has many difficulties in acting against new URLs, so we are increasingly focused on machine learning algorithms, and that is precisely the focus of this paper. In this paper, we'll cover the most common and dangerous attacks through malicious URLs, how they work, and how to prevent them. We will focus on and analyze more concretely a technique to detect, which uses algorithms of Machine Learning.

Keywords: Malicious URLs, Machine Learning, Detection, Algorithms.

1 Introduction

With the rise of the internet and the evolution of technologies, we have moved away from the society of industries and we have moved on to the information society. We are in a society where we use various forms and technologies of communication and information, such as online purchases, online banking and betting sites. But with the increase of the use of these technologies, also comes the increase of risks, because third parties try to steal and take advantage of our information, networks and systems for their own benefit. [1]

People using, for example, home computers are targets that are very vulnerable to external attacks and threats because in most cases they are not aware of the various types of attacks and how they should be protected. They are naiver and fall into the simplest attacks. Even people who use computers at work and are alerted and aware of the attacks and that they must be careful and protect the company and themselves fall into the simplest attacks. [1]

These attacks can be reconnaissance only, which means that attackers attempt to map systems and networks in order to find flaws, can be access attacks, where attackers gain access to systems and networks, gain access to services, databases, among others. Another form of attack, is to make a machine or network resource unavailable or make it so slow that it is practically impossible to use it. [2]

Going to the numbers and the facts, 24,000 malicious applications are blocked every day, and information that most applications release is 63% mobile phone numbers and 37% device location. From 2016 to 2017 the percentage of cybersecurity costs increased by 22.7%, with malware attack costing companies 2.4 million dollars on average. The applications with most

security problems are those of lifestyle and music, 411 million accounts of dating sites were stolen in 2017. [3] [4] Cybersecurity has emerged as a set of tools, policies, concepts, protocols, actions, techniques, and best practices that help protect organizations and individuals in general from attacks, helping to maintain confidentiality, integrity, authenticity and nonrepudiation of information. [5]

2 Universal Resource Locator

URL's, also known as Universal Resource Locator, as the name indicates are used to find a particular resource on the Internet, they are also known as web address. A URL finds by providing, to the browser for example, an abstract of the location of the resource, when this resource is found the system can execute a great diversity of operations. [6]

A URL is formed by the protocol used to access the resource, the location of the server to be accessed, which may be on the form of the domain or on the form of the IP address and the path where the resource is located. [7]

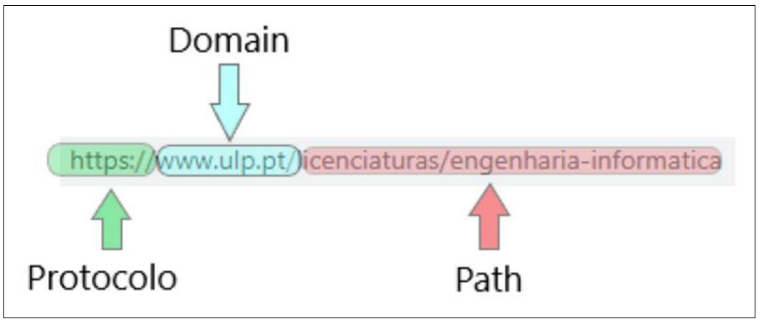

Fig.1. This figure, represents an URL and consequently all the parts that constitute an URL.

The following protocols are the most commonly used today: [8]

- FTP, File Transfer Protocol, allows the transfer of files between two machines. The client connects to the server to obtain some file and the server receives the request and supplies the file.
- SFTP, Simple File Transfer Protocol, does the same as the FTP protocol, but uses a different technology that allows you to authenticate and secure the connection between the client and the server, ensuring greater file security.
- POP3, Post Office Protocol, works like a mailbox for emails, when the client accesses the server, has access to all the emails that he received.

- SMTP, Simple Mail Transfer Protocol, this protocol is aimed at sending emails, this protocol is very effective, but only allows the sending of text.

- IMAP, Internet Message Access Protocol, this protocol joins the best of the two worlds, allowing the user to access and manage their messages directly on the server, in a counterpart you have to always be connected to the network and it has storage limit.

- HTTP, Hypertext Transfer Protocol, it is used for site navigation. The protocol also works with a connection between the client and the server, where the client is the browser that the user uses, and the server is the site that is intended to be accessed. The browser sends a request to access a page and the server sends an access permission response. With it, come the files that form the page that the user wants to access.

- HTTPS, Hypertext Transfer Secure, works like HTTP but has a layer of protection because it uses SSL, the SSL protocol ensures privacy and data integrity between two applications that communicate over the internet. This is done through the authentication of the parties involved and the encryption of the data transmitted between the parties.

When a URL directs the browser to a file that it can open, such as images or PDF's, the browser displays content without having to download the file, but many other types of files require a download. Because of all the complexity and diversity of functions, URLs can also be made to do evil and attack the user. [9]

3 Attacks using URL

One of the most common attacks is called phishing, the main purpose of the attack is to trick the user into giving them their data, usually login data. The attack consists of getting the target to click on a link that takes you to a page similar to what the target wanted, but when they login, the attackers get their login data. [10]There are several techniques for phishing, one of them is to buy a domain similar to the site that we want to copy for example facebok.com instead of facebook.com, because the human brain is accustomed to when reading words with almost imperceptible errors, it processes as if the error did not exist, then the user does not even notice that he is not on the real site and logs in giving access to his data to the attackers. In an experiment at Carleton University, giving a list of links to multiple users to identify if it is an attack attempt or the original site, users used various techniques to verify that the site was true or false by analyzing the URL, testing the features of the sites, looking at google if the URL of the original site and the one provided for the experience were the same, checking if the site uses SSL, among others, but outside of this experiment, people do not do this for all sites that enter , only if they notice that some is suspect. The problem is that they try hard to go unnoticed. [11] Another type of attack is the Driven-by download, which consists in unintentionally downloading malicious code, this attack does not require the user to click something or do

something. The purpose of these codes is to communicate with another computer to gain access to the device. These malicious files are very common in corrupted websites. [12]

There are attacks that are based more on the psychological of the person, such as social engineering, that the first step of this attack is to know your target, such as sites they visit the most, their concern about security, their location, among other data. This attack is very common because it is easier to identify the vulnerabilities of people than of systems and networks. [13] Some attacks using this technique are [14]:

- Bating, these attacks try to wake greed and curiosity to the victim, they can be physical like leaving a PEN in a strategic place so that when the victim finds it, it enters in your computer and installs malware in your system, or through ads that lead to malicious sites that lead to malware downloads.

- Scareware, victims receive false alarms and threats leading them to believe that their system is infected with malware. Usually are websites that say these messages and almost always have a button to download a tool that will correct all the evils of the system, but ironically this tool will infect the system.

- Honey trap, gaining the confidence of the target, until you are in a kind of online relationship, with the confidence of the target, you can send a malicious URL that the person will most likely trust you and provide their data.

3.1 Malicious URL Detection using Backlists

To combat these malicious URLs, we must detect them because of their ease of deceiving us, so we created forms to detect them, one of the most common being the use of Blacklists. Blacklists are databases where the data of the URLs that have already been confirmed as malicious are saved, and more URLs are added to the list over time. Whenever a new URL is visited, a database lookup is performed. If the URL is present in the blacklist, it is considered to be malicious and then a warning will be generated; else it is assumed to be benign. Although it seems a safe and effective method, backlists are slow, because they cannot keep up with the growing number of URLs, meaning that these databases will never be able to have all the malicious URLs that exist because new ones are created every day and new ways to get around blacklists. To combat this problem and find a new way to detect malicious URLs, scientists have, in recent years, sought a solution in Machine Learning algorithms. [15]

4 Malicious URL Detection using Machine Learning.

Machine learning is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. Machine Learning

approaches, use a set of URLs as training data, which sometimes can be Blacklists, and based on the statistical properties, learn a prediction function to classify a URL as malicious or benign, which gives them the ability to generalize to new URLs unlike blacklisting methods. [16]

4.1 Extraction of Features

The first requirement for a Machine Learning model is a "training date," which corresponds to a large number of URLs. These algorithms can be supervised or unsupervised, meaning that the algorithm knows whether URLs are malicious / benign or do not know. There are also semi-supervised algorithms which mean that they know the classification and a part of the training URLs. The next step is to extract information from the URL and transform it so that it can be rendered by the template. [17]

These features can be: [18]

- Lexical, are features obtained through the properties of the URL, because through the aspect of the URL it should be possible to identify if it is malicious, because these URLs try to pretend to be benign URLs, changing its aspect a bit. The most commonly used features are the length of the URL, the length of each URL component (Domains, subdomains, path), number of special characters, and each character sequence separated by a special character ("/", ". ") Is considered a feature. Based on all the words in all the URLs, a dictionary was built. If the word was present in the URL, the value of the resource would be 1 and 0 otherwise. This is also known as the bag-of-words model. The whole bag-of-word resource approach can be seen as a form of blacklist compatible with Machine Learning, instead of focusing on the entire URL string, analyze the URL based on smaller components. Not to be detected by Blacklists hackers use algorithms to generate URLs that are not in them but it will be harder to go through the bag-of-words approach because the template will detect that the URL has words that are not in the dictionary. • Host-based, are obtained through the host properties of the URL. This allows us to know the location of the host, its identification and some other features. URL lifetime is one of the most important features since malicious users typically have much less lifetime than benign ones. It is possible to obtain information about when and who registered the domain as well as the location of the IP. Due to the difficulty of obtaining new IPs these features are very important to detect malicious URLs.

- Content Based, these features are obtained by downloading the contents of the URL page, thus being the most dangerous type of feature to obtain but can greatly help prediction accuracy. We can get the HTML code of the page and analyze the number of words, average words per line, distinct words, links to remote scripts and invisible objects. Often, malicious code is encrypted in HTML and this is directly linked with a larger number of words with great length and with the use of concatenation. Other

features we can get are related to JavaScript as they are used by hackers to encrypt malicious code or execute without permission. We analyze features such as the number of long strings, number of events, number of strings, suspicious objects and tags, and the use of functions such as exec (), link (), eval (), escape (), search () because they are used to distribute malware. [19]

- Other, there has currently been a growth of platforms that allow you to shorten a URL to a totally different and smaller string to allow sharing on social networks, such as Twitter that, per tweet, only allows 140 characters. But these platforms can be used for evil and spread malicious URLs, although they try not to produce short URLs for the malicious, they have little effectiveness since they use Blacklists as a basis to block the malicious ones. So, to bridge this, we can get information about the shortened link itself and the devices that accessed the platform to shorten them as well as the devices that shared them, also counting the number of shares on social networks. Other features that may help in detection are the measures of popularity, one of these measure is the Link Popularity which is scored based on incoming links from other webpages. Also used are the number of popups and the behavior of plugins. The number of times that the URL is accessed can also be an indicator, because malicious ones, are less accessed. [20]

The next step is to transform and convert all the features extracted from the URLs into a numerical vector so that they can be powered by machine learning algorithms. In addition, some data normalizations can usually be used to deal with the scale problem. [17]

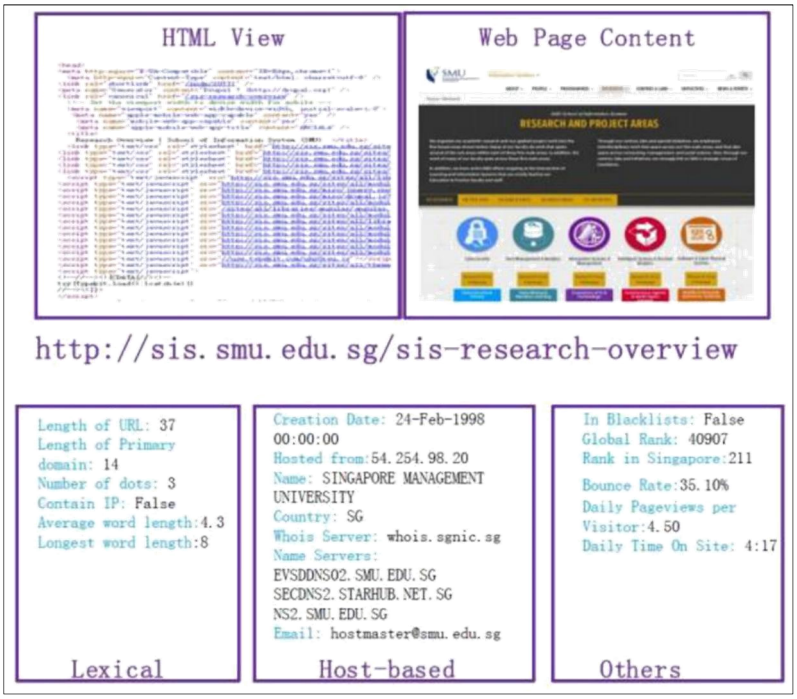

Fig.2. This figure represents the features that can be obtain from an URL.

4.2 Machine Learning Algorithms for Malicious URL Detection

After converting the features to vectors, there are many algorithms that can be applied to predict if an URLs is malicious or not, so we are going to study some of them. We are going to study some algorithms from the Batch Learning family and others from the Online Learning family. The Batch Learning algorithms, work with the assumption that all the training data is available after the training task, some of the Batch Algorithms are:

- Naive Bayes, this algorithm generatively classifies the URLs and it's a "naive" algorithm, in the sense that it considers that all the features (x) are independents from each other and, for each feature, it calculates the conditional probability, described in the next equation, are y=1 stands for URL malicious and y=0 for URL benign. After making this calculation for all features it decides if its malicious or benign. [22]

$$P(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=1) + P(\mathbf{x}|y=0)}.$$

- Support Vector Machine (SVM) are well-known for binary classification of great dimensions of data. This algorithm uses a single rule that is expressed by kernel function K(x, x ′) that computes the similarity between two feature vectors and non-negative coefficients {αi} n i=1 that indicate which training examples lie close to the decision boundary. SVMs classify new examples by computing their (signed) distance to the decision boundary. Up to a constant, this distance is given by

$$h(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i (2y_i - 1) K(\mathbf{x}_i, \mathbf{x}),$$

  The sign of this distance indicates the side of the decision boundary on which the example lies. In practice, the value of h(x) is to predict a binary label for the feature vector x. [21]

Online Learning are algorithms that treat the data as an instance flow, thus they are Learning from the training data and predicting the real URLs almost at the same time. So, from the Online Learning family, here are some of them:

- First order, are algorithms that learn by updating a vector with the labels (benign or malicious) using only first order features and the training data. [23]

- Second order, these algorithms, instead of using first order features, tries to boost the learning efficiency, by exploring second order features like statistical features. [23]

- Online Active Learning, supervised algorithms assume that the classifications of the training data are received by them with no cost, but that's not true, in real systems this

process can be expensive and slow, to overcome this issue Online Active Learning tries to reduce this cost. These algorithms aim to only consult if an URL was already classified (malicious or benign) if it low transmits confidence levels. [24]

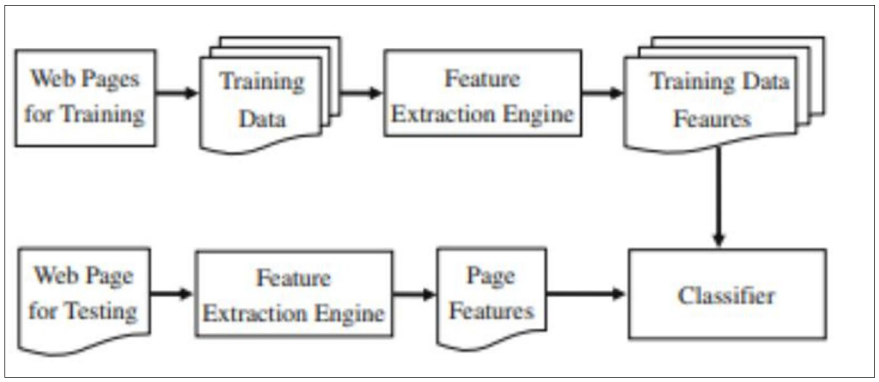

Fig.3. This figure represents all malicious URL Detectors, based on Machine Learning.

5 Conclusion

In this paper we learn that through some techniques, hackers can obtain our personal and private information and that they can bypass several of our ways to detected them and protect from them. One of these serious treats and the main focus of this paper was the malicious URL, hackers have several techniques and algorithms to obfuscate their URLs to bypass out defenses. To overtake them in this race to protect our data, one promising choice, that we aboard on the paper, are the Machine Algorithms projected to Detect or classify URLs as benign or malicious. Although they are a good way to improve security they are expensive and hard to adapt to some applications, like browsers, but because of their potential, there is a need to wager in them and to study more, so that they can grow and be more present in people's life.

References

1. Wang, W.; Lu, Z.: Cyber security in the Smart Grid: Survey and Challenges. Computer Networks, vol. 57:5, pp. 1344-1371. ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2012.12.017. (2013)
2. Uma M., Padmavathi G.: A Survey on Various Cyber Attacks and Their Classification, International Journal of Network Security, vol.15:5, PP.390-396. (2013)
3. Internet Security Threat Report, Vol.23, Symantec.
4. Richards, K.; LaSalle, R.; INSIGHTS ON THE SECURITY INVESTMENTS THAT MAKE A DIFFERENCE, COST OF CYBER CRIME STUDY, Accenture. (2017)
5. Rossouw, S.; Niekerk, J.;From information security to cyber security, Computers & Security, vol. 38, pp. 97-102, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2013.04.004.(2013)
6. What is URL?, https://searchnetworking.techtarget.com/definition/URL, last accessed 22/11/2018.
7. What is URL? Definition from Techopedia, https://www.techopedia.com/definition/1352/uniform-resource-locator-url, last accessed 22/11/2018.
8. Protocolos de Rede, https://www.weblink.com.br/blog/tecnologia/conheca-os-principaisprotocolos-de-internet/ last accessed 24/11/2018.
9. What's the difference between a site and a URL?, https://www.lifewire.com/what-is-a-url2626035, last accessed 24/11/2018.

10. Ataque Homografico, https://www.techtudo.com.br/noticias/2017/11/ataque-homograficotruque-na-url-engana-usuarios-com-paginas-falsas.ghtml, last acessed 30/11/2018.
11. Alsharnouby, M.; Alaca, F.; Chiasson, S.; Why phishing still works: User strategies for combating phishing attacks, International Journal of Human-Computer Studies, Volume 82, pp. 69-82, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2015.05.005.(2015)
12. What is a Drive-by Download, https://www.kaspersky.com/resource-center/definitions/drive-by-download, last acessed 30/11/2018.
13. What is Social Engineering?, https://searchsecurity.techtarget.com/definition/social-engineering, last acessed 30/11/2018.
14. What is Social Engineering?, https://www.incapsula.com/web-application-security/socialengineering-attack.html, last acessed 1/12/2018.
15. Sahoo, D.; Liu, C.; Steven C. H.; Malicious URL Detection using Machine Learning: A Survey, eprint arXiv:1701.07179. (2017)
16. What is machine learning?,https://searchenterpriseai.techtarget.com/definition/machinelearning-ML, last accessed 12/12/2018.
17. Detecting Malicious URLs with Machine Learning, https://ritcsec.wordpress.com/2017/12/07/detecting-malicious-urls-with-machine-learning/, last accessed 11/12/2018.
18. Utilizando Aprendizado de Máquina para Detecção Automática de URLs Maliciosas Brasileiras, Lucas Dantas Gama Ayres, Salvador – BA. (2018)
19. Y.-T. Hou, Y. Chang, T. Chen, C.-S. Laih, and C.-M. Chen, "Malicious web content detection by machine learning," Expert Systems with Applications, vol. 37:1, pp. 55–60. (2010)
20. Sahoo, Doyen; Liu, Chenghao; Hoi, Steven C. H, "Malicious URL Detection using Machine Learning: A Survey", arXiv e-prints, January 2017.1
21. Kolari, P.; Finin, T.; Joshi, A.; "Svms for the blogosphere: Blog identification and splog detection," in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 92–99. (2006)
22. Aggarwal, A.; Rajadesingan, A.; Kumaraguru, P.; "Phishari: automatic realtime phishing detection on twitter," in eCrime Researchers Summit (eCrime), IEEE, pp. 1–12. (2012)
23. Ma, J.; Saul, L.; Savage, S.; Voelker, G. M.; "Identifying suspicious urls: an application of large-scale online learning," in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 681–688. (2009)
24. Zhao, P.; Hoi, S. C.; "Cost-sensitive online active learning with application to malicious url detection," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 919–927. (2013)