Phishing URL Detection

Bar Goldman, Kirill Perevalov
Ariel University

Abstract

Phishing is a type of cyber attack where someone impersonates a legitimate website page in order to collect sensitive information from the victim and use it maliciously. To combat the risks associated with this threat, it is important to identify phishing sites in time. Machine learning models work well for this purpose since they can predict phishing incidents using basic information about the sites. In most of the most modern solutions dealing with phishing detection, several techniques for identifying phishing have been proposed, such as third-party and heuristic approaches. These approaches suffer from a number of limitations that need to be addressed in order to identify phishing addresses. Therefore, this article will discuss our contribution to the identification of phishing sites by proposing a machine learning model that identifies phishing URL addresses using lexically based features, which do not rely on any third-party services and do not require visiting the web page for feature extraction.

1 Introduction

Phishing is a cybercrime attack that is persistently and dangerously escalating. A phishing attack can harm any type of business, whether governmental, social, financial, or personal. It is a type of attack where someone impersonates a legitimate website page in order to collect sensitive information from the victim such as email, passwords, phone numbers, and credit card details and use them maliciously. Attackers need to create websites that look very similar to their legitimate counterparts they are trying to impersonate and spread the links through fake email messages or other communication channels.

As stated in the Phishing Activity Trends Report [1]. published by the Anti-Phishing Working Group (APWG) in November 2023, in the first quarter of 2023, APWG observed 1,624,144 phishing attacks. This is a record observation - the worst quarter for phishing that APWG has ever observed. The total increased from 888,585 in the fourth quarter of 2022, and above 1,270,883 phishing attacks in the third quarter of 2022, which was the record at that time.

	January	February	March
Number of unique phishing Web sites (attacks) detected	495,690	509,394	619,060
Unique phishing email campaigns	40,863	45,259	40,742
Number of brands targeted by phishing campaigns	561	549	576

Figure 1:

To protect against the risks arising from phishing attacks, the identification of phishing sites is of utmost importance. Distinguishing between genuine web pages and phishing pages, when they look identical, is difficult for human eyes. The only thing that separates these web pages from each other is the URL address. A URL can be distinguished and easily identified as a phishing URL successfully, making machine learning-based approaches particularly suitable for classifying sites as phishing or legitimate based on their addresses.

2 Related Work

The identification of phishing sites is a crucial area that has garnered significant interest from many researchers. There are two types of identification methods, which can be classified into the following categories: third party based detection and heuristic based detection.

Third-party Based Detection Most research work relies on third-party-based approaches, namely blacklists and whitelists. They focus on creating and maintaining lists of phishing and legitimate sites, and the most up-to-date browsers, such as Google Chrome, Microsoft Edge, Opera, and Mozilla, keep a list of forbidden and allowed URLs. The problem is that these approaches fail when encountering zero-day phishing sites. [2, 3, 4].

Heuristic Based Detection Heuristic detection divides into two main strategies:

Textual and Visual Similarity: It examines the textual and visual similarity between dubious web pages and legitimate ones that may be imitations in phishing schemes. In these cases, the imitator can change a very small part of the web page without even changing the content of the web page, thus not being caught as a phishing site.

URL-based Techniques: The URL-based method relies only on the characteristics of the website address and is divided into two parts: lexical and host-based. Lexical-based features are extracted from the address characteristics and strings, while host-based features are extracted from the WHOIS database.

Among all these techniques, URL (lexical-based) is the only one that does not depend on the internet, its processing time is very low, and the accuracy of identifying phishing URLs is very high.

In this article, we chose to focus on URL-based techniques, wishing to smartly identify a phishing site with the help of a machine learning mechanism, which will know how to identify these sites in advance and alert about them.

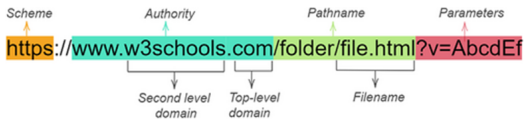

Figure 2: URL structure

As shown in Figure 2, a URL consists of several components, namely,

- **Scheme:** , which specifies the protocol used for making the request.
- **Authority:** which specifies the Fully Qualified Domain Name of the server hosting the website and its two main components, that is, second level domain and top-level domain.
- **Pathname:** of the resource being requested, that includes the folder as well as the filename.

- **Optional parameters:** parameters corresponding, for example, to a query.

Although all website addresses share the same structure, the addresses of phishing sites usually differ from those of legitimate sites. In fact, attackers create their website addresses to make them look as similar as possible to legitimate addresses. Therefore, the feature extraction process needs to take into account these differences and, in particular, the tactics and strategies implemented by attackers to deceive people. Based on information we collected, we found that phishers use various techniques to obscure the true identity of the site in different components:

Domain: Used in subdomains For example, 'login.examplebank.com.fakebank.com' may appear as a login page for "examplebank," but in fact, it is hosted under "fakebank.com".

Misspelling of domain names Phishers write domain names with spelling errors, they take pre-known, familiar domains and make a slight change in the spelling of domain names For example, 'www.googlle.com' instead of 'www.google.com'.

Path: Embedding the real domain name in the path The actual domain name may be embedded somewhere in the path of the website address, misleading users to focus on the legitimate part while ignoring the rest of the website address. For example, 'http://www.fakebank.com/www.examplebank.com/login'.

Length of website addresses: Phishers use URL shortening services that obscure the final destination of the link, which does not allow determining the legitimacy of the website address without clicking on it.

In our article, we decided to extract features from every part of the website address and give a ratio to each component separately (including the entire URL), then we used the random forest technique for extracting relevant features. We saw that most existing articles extract features from the full URL address like the length of the addresses, various symbols appearing in it, etc... In our article, we chose to focus on extracting features from the full URL address as well, but also to emphasize each component individually.

3 The contributions achieved

We proposed a feature-rich framework, initially extracting 79 features based on the entire URL address as well as from the full domain name, path name, file name, and parameters. And 12 leading features were selected using the Random Forest technique.

4 Evaluation

4.1 Dataset Description

To test the proposed approach, real data pertaining to phishing website addresses and legitimate sites were collected from 3 different sources: PhishTank [5], OpenPhish [6] for phishing sites, and for legitimate sites, the Alexa top 1 million [7].

We compiled a balanced dataset consisting of 45,727 website addresses referring to phishing sites and 45,726 website addresses referring to legitimate sites.

4.2 Feature Extraction

During our research, we created 79 features based on the entire URL address as well as from the full domain name, path name, file name, and parameters.[8]

Examination of the Number of Symbols These features relate to the appearance of 17 different symbols within the complete website addresses and identified components. In fact, an excessive number of symbols is a strong indicator of a phishing website address.

Length These features take into account the length, namely the number of characters of the entire website address and its individual components. Generally, attackers tend to use long URLs to confuse people and make them believe the website address is legitimate.

IP Address in Domain This is a binary feature used to indicate the presence of an IP address instead of the full domain name within a URL.

Additional features refer to the number of hyphens in the domain name, the presence of an email address in the website address, the presence of words such as "server" or "client" in the domain name, and the number of parameters.

4.3 Feature Selection

The most relevant features are selected from those extracted by the Random Forest technique. In particular, features capable of distinguishing between phishing sites and legitimate sites are preferred and selected. At the end of this stage, 12 features are considered the most important:

Feature	Description
num_._url	Number of periods in the URL
num_-_url	Number of hyphens in the URL
num_/_url	Number of slashes in the URL
length_url	Length of the URL
num_dots_dom	Number of periods in the domain
num_hyph_dom	Number of hyphens in the domain
num_vowels_dom	Number of vowels in the domain
length_dom	Length of the domain
num_subdomains	Number of subdomains
num_slash_path	Number of slashes in the path
length_path	Length of the path
length_file	Length of the file

Table 1: Description of URL Features

5 Results

In this section we will review the ML results:

Table 2: Classification Report

	Precision	**Recall**	**F1-score**	**Support**
0	0.98341	0.99792	0.99061	9145
1	0.99789	0.98316	0.99047	9146
Accuracy			0.99054	18291
Macro Avg	0.99065	0.99054	0.99054	18291
Weighted Avg	0.99065	0.99054	0.99054	18291

6 Summary

In this study, a machine learning model for identifying phishing URL addresses was presented, based on a dataset containing both malicious and non-malicious URLs. Various features based on the entire URL address as well as from the full domain name, path name, file name, and parameters were introduced. These features do not rely on any third-party services and do not require visiting the web page for feature extraction. The feature extraction uses the Random Forest technique. The selected features will enable us to train our machine learning model to classify each website address in the best way possible.

References

[1] Anti-Phishing Working Group (APWG). *Phishing Activity Trends Report*, 2023. Available online; accessed 20 February 2024.

[2] Author's Name. "Adopting automated whitelist approach for detecting phishing attacks," *ScienceDirect*, Year, Volume(Issue Number), Page Numbers. Available online; accessed 20 February 2024.

[3] Author's Name. "Anti-phishing based on automated individual whitelist," *Proceedings of the 4th ACM workshop on Digital identity management*, ACM, Year, Page Numbers.

[4] Author's Name. "A novel approach to protect against phishing attacks at client side using auto-updated white-list," *EURASIP Journal on Information Security*, Springer, Year, Volume(Issue Number), Page Numbers. Available online; accessed 20 February 2024.

[5] PhishTank. URL: `https://phishtank.org/`, Accessed: 20 February 2024.

[6] OpenPhish. URL: `https://openphish.com/`, Accessed 20 February 2024.

[7] Kaggle Dataset, "Top 1 Million Domains," Available at `https://www.kaggle.com/datasets/cheedcheed/top1m/data`, Accessed 20 February 2024.

[8] Authors. "Title of the article," *Journal Name*, Volume(Issue), Year, Pages. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920313202, Accessed 20 February 2024.