DATA NETWORKS LAB

PHISHING URL DETECTION WITH MACHINE LEARNING

TANUMON ROY

174259, ¾ ECE-B, NIT WARANGAL

YASH DESHMUKH

174269, ¾ ECE-B, NIT WARANGAL

Introduction

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels.

Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or

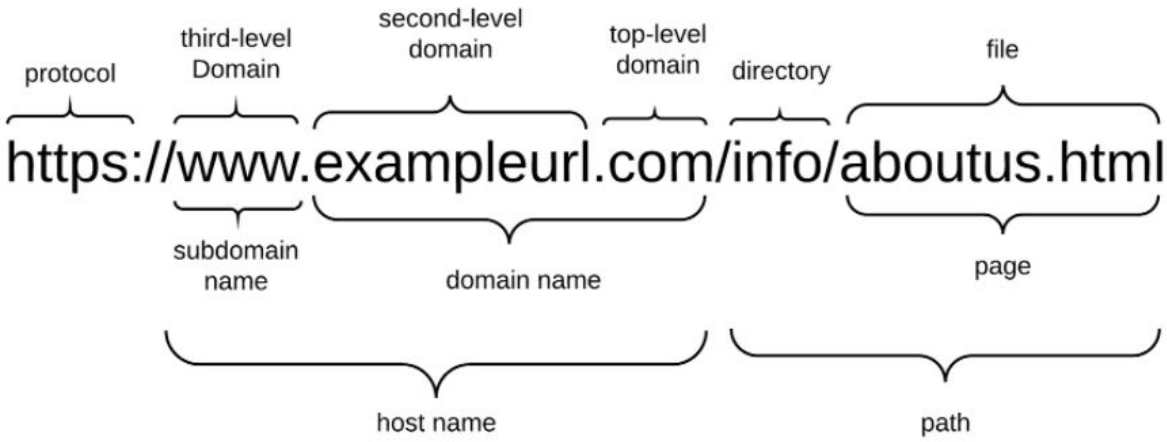
has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.

Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents.

Characteristics of Phishing Domains

Let's check the URL structure for the clear understanding of how attackers think when they create a phishing domain.

Uniform Resource Locator (URL) is created to address web pages. The figure below shows relevant parts in the structure of a typical URL.



It begins with a protocol used to access the page. The fully qualified domain name identifies the server who hosts the web page. It consists of a registered domain name (second-level domain) and suffix which we refer to as top-level domain (TLD). The domain

name portion is constrained since it has to be registered with a domain name Registrar. A Host name consists of a subdomain name and a domain name.

A phisher has full control over the subdomain portions and can set any value to it. The URL may also have a path and file components which, too, can be changed by the phisher at will. The subdomain name and path are fully controllable by the phisher. We use the term FreeURL to refer to those parts of the URL in the rest of the article.

The attacker can register any domain name that has not been registered before. This part of the URL can be set only once. The phisher can change FreeURL at any time to create a new URL. The reason security defenders struggle to detect phishing domains is because of the unique part of the website domain (the FreeURL). When a domain is detected as fraudulent, it is easy to prevent this domain before an user accesses it.

The most common method to detect malicious URLs deployed by many antivirus groups is the blacklist method. Blacklists are essentially a database of URLs that have been confirmed to be malicious in the past. This database is compiled over time (often through crowd-sourcing solutions, e.g. PhishTank), as and when it becomes known that a URL is malicious. Such a technique is extremely fast due to a simple query overhead, and hence is very easy to implement.

Additionally, such a technique would (intuitively) have a very low false-positive rate (although, it was reported that often blacklisting suffered from non-trivial false-positive rates). However, it is almost impossible to maintain an exhaustive list of malicious URLs, especially since new URLs are generated everyday. Attackers use creative techniques to evade blacklists and fool users by modifying the URL to “appear" legitimate via obfuscation.

To overcome these issues, in the last decade, researchers have applied machine learning techniques for Malicious URL Detection. Machine Learning approaches use a set of URLs as training data, and based on the statistical properties, learn a prediction function to classify a URL as malicious or benign. This gives them the ability to generalize to new URLs unlike blacklisting methods. The primary requirement for training a machine learning model is the presence of training data. In the context of malicious URL detection, this would correspond to a set of large numbers of URLs.

There is a rich family of machine learning algorithms in literature, which can be applied for solving

malicious URL detection. After converting URLs into feature vectors, many of these learning

algorithms can be generally applied to train a predictive model in a fairly straightforward manner.

Features Used for Phishing Domain Detection

There are a lot of algorithms and a wide variety of data types for phishing detection in the academic literature and commercial products. A phishing URL and the corresponding page have several features which can be differentiated from a malicious URL. For example; an attacker can register a long and confusing domain to hide the actual domain name **(**Cybersquatting, Typosquatting). In some cases attackers can use direct IP addresses instead of using the domain name. This type of event is out of our scope, but it can be used for the same purpose. Attackers can also use short domain names which are irrelevant to legitimate brand names and don't have any FreeUrl addition. But these types of web sites are also out of our scope, because they are more relevant to fraudulent domains instead of phishing domains.

Beside URL-Based Features, different kinds of features which are used in machine learning algorithms in the detection process of academic studies are used. Features collected from academic studies for the phishing domain detection with machine learning techniques are grouped as given below.

1. URL-Based Features
2. Domain-Based Features
3. Page-Based Features
4. Content-Based Features

URL-Based Features

URL is the first thing to analyse a website to decide whether it is a phishing or not. As we mentioned before, URLs of phishing domains have some distinctive points. Features which

are related to these points are obtained when the URL is processed. Some of URL-Based Features are given below.

- Digit count in the URL
- Total length of URL
- Checking whether the URL has been Typosquatted or not. (google.com → goggle.com)
- Checking whether it includes a legitimate brand name or not (apple-icloud-login.com)
- Number of subdomains in URL
- Is Top Level Domain (TLD) one of the commonly used one?

Domain-Based Features

The purpose of Phishing Domain Detection is detecting phishing domain names. Therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us. Some useful Domain-Based Features are given below.

- Its domain name or its IP address in blacklists of well-known reputation services?
- How many days passed since the domain was registered?
- Is the registrant name hidden?

Page-Based Features

Page-Based Features are using information about pages which are calculated reputation ranking services. Some of these features give information about how reliable a website is. Some of Page-Based Features are given below.

- Global Pagerank
- Country Pagerank
- Position at the Alexa Top 1 Million Site

Some Page-Based Features give us information about user activity on the target site. Some of these features are given below. Obtaining these types of features is not easy. There are some paid services for obtaining these types of features.

- Estimated Number of Visits for the domain on a daily, weekly, or monthly basis
- Average Pageviews per visit
- Average Visit Duration
- Web traffic share per country
- Count of reference from Social Networks to the given domain
- Category of the domain
- Similar websites etc.

Content-Based Features

Obtaining these types of features requires active scan to the target domain. Page contents are processed for us to detect whether target domain is used for phishing or not. Some processed information about pages are given below.

- Page Titles
- Meta Tags
- Hidden Text
- Text in the Body
- Images etc.

By analysing these information, we can gather information such as;

- Is it required to login to website
- Website category
- Information about audience profile etc.

Detection of Phishing Domains : Our Perspective

Procedure

In this project we have taken the following steps to build a Machine Learning model which can identify phishing domains.

1. **Collect URL data** from various surveys and websites

   We collected data from various surveys and websites like:

   Alexa-Top 1M websites (for benevolent URLs) [https://www.alexa.com/topsites]

   Forbes-Global 2000 companies (for benevolent URL subdomain names) [https://www.forbes.com/global2000/]

   PhishTank (for malicious URLs) [https://www.phishtank.com/developer_info.php]

2. **Extract the features** from the URLs and create datasets for both phishing and benevolent URLs

   We defined a python function which takes in a URL as input and returns its various features in the form of a pandas DataFrame. Then we took URLs from the collected data, passed them through the function, and formed datasets of our own with the URL features.

3. **Mix and shuffle** the phishing and benevolent URL datasets and **split** the mixed dataset into training and testing datasets

   We mixed all the feature-extracted data together and did a 80:20 split on it.

4. **Create a Neural Network model** with a particular set of hyperparameters

   We defined a NN model which had four layers (input layer with number of features [n] units; hidden layer with n units; hidden layer with 2n units; output layer with 1 unit).

5. **Train the model** with the training data

We trained the NN model over the training data with 10 epochs (iterations)

6. **Test the model** with the testing data
7. Observe the loss and accuracy and **improve the model** hyperparameters accordingly
8. **Repeat** steps 4 through 7 with new model configurations

Result

After training the NN model and testing it, we achieved an accuracy of about 79%.

This can be improved more by introducing more data and features because the model has a slight tendency of underfitting with the given feature set and the accuracy does not seem to improve much with more number of epochs.

Acknowledgment

https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5

Malicious URL Detection using Machine Learning: A Survey [https://arxiv.org/pdf/1701.07179.pdf]

DOYEN SAHOO, CHENGHAO LIU, STEVEN C.H. HOI,

School of Information Systems, Singapore Management University