Network Security Project Report

Group Members:

Muhammad Maaz Siddiqui
Umar Khalid
Sarah Ebrahim
Majeed Toorie

Institution:
Institute of Business Administration Karachi

Teacher:
Faisal Iradat

Abstract

Malicious websites pose a significant threat, requiring the analysis of numerous URLs and the development of comprehensive blacklists. This report presents the first version of a dataset derived from a web security project conducted by our group as part of our academic curriculum. The project aims to fill the gap in available datasets and improve results through ongoing work.

Content

The project involved evaluating various classification models to predict malicious and benign websites based on application layer and network characteristics. The dataset was obtained from verified sources of benign and malicious URLs, using a low interactive client honeypot to isolate network traffic. Additional information, such as server country obtained through Whois, was also collected.

URL Dataset

The dataset for this project is derived from MalCrawler [1], a web crawler designed for seeking and crawling malicious websites. Labels for the dataset have been verified using the Google Safe Browsing API [2]. The attributes in the dataset include extracted features from websites, and the raw page content, including JavaScript code, allowing for the utilization of unstructured data in Deep Learning or for extracting further attributes. The selection of attributes is based on their relevance [3].

References:

[1] Singh, A. K., and Navneet Goyal. "MalCrawler: A crawler for seeking and crawling malicious websites." In International Conference on Distributed Computing and Internet Technology, pp. 210-223. Springer, Cham, 2017.
[2] Google Safe Browsing API
[3] Singh, A. K., and Navneet Goyal. "A Comparison of Machine Learning Attributes for Detecting Malicious Websites." In 2019 11th International Conference on Communication Systems & Networks (COMSNETS), pp. 352-358. IEEE, 2019.

Framework

Python scripts were developed for systematic URL analysis, filtering the dataset to 63,191 URLs. The scripts will be made available to the open-source community on GitHub.

Feature Generator

Dynamic and static analyses were employed to study malicious websites, focusing on features from the application layer and network layer.

Data Description

The dataset includes various features such as URL length, the number of special characters, character set, server information, content length, Whois details, TCP conversation exchange, remote TCP ports, remote IPs, bytes transferred, and various packet-related metrics. The 'TYPE' variable categorizes web pages as 1 for malicious and 0 for benign.

Conclusions and Future Works

The report provides initial insights into the dataset and its creation process. Ongoing work includes improving classification models and expanding the dataset.

Acknowledgements

We would like to express our gratitude to our institution and professor for their guidance and support throughout this project. If the dataset is used in research, we kindly request citing our paper (Urcuqui et al., 2017).