Malicious Domain Detection from External Data Sources

Bc. Adam Horák

supervisor Ing. Radek Hranický, Ph.D.

Implementing a Domain Classifier

1. Gathering Data

Name lists from reputable sources for both benign and malign domains are gathered and as much relevant data as possible is gathered for each of them to offer maximum flexibility when extracting distinct features.



2. Feature Extraction & Engineering

From Mongo documents to a table of features using a custom automated data transformation pipeline. To find out what combination of things gives away each of the malicious domain category, various features are extracted from the rich data we collected.

	FEATURE IMPORTANCE
PHISHING	time before leaf certificate expires
MALWARE	time before root certificate expires
DGA	domain registration date
CRYPTO	subject alternative name extension name count
TRACKERS	DNS TXT record count
one	time before domain expires

3. Model Training & Evaluation

The extracted features are then evaluated for importance by training a gradient boosting model to get their gain scores. Iterating this process helps to identify the most informative features for accurately categorizing malicious domains.

► XGBOOST ► IMPORTANCE DATASET

Deploying to Protect a Network

NETWORK

DOMAIN

Domains from network traffic monitoring are examined on demand by the classifier to predict threats.

DOMAIN CLASSIFIER

DATA RESOLVER Pulls data from external sources for the examined domain.

TRAINED MODEL

Predicts if the domain might belong to one of our malicious classes based on the resolved data.

SAFETY REPORT

95% OK

Network administrators can validate and respond to malign domains.

90% PHISHING

50% MALWARE

99% DGA



Using the same ensemble

SOURCES

DOOFENSHMIRTZ INC.

In a bid to make some quick cash, Dr. Heinz Doofenshmirtz decides to set up a phishing scheme online.

PHISHING DOMAIN SETUP

He starts by registering his domain and adding it to the Domain Name System. We can use this data against him.

PLANS FOILED BY OUR CLASSIFIER

His plans are quickly foiled when our predictive domain classifier accurately identifies and blocks his malicious domain. By analyzing the domain's features and its registration data, our technology was able to make a prediction and prevent his intended victims from falling for his scheme.

The research is supported by "Flow-based Encrypted Traffic Analysis" project, no. VJ02010024 granted by Ministry of the Interior of the Czech Republic and the "Smart information technology for a resilient society" project, no. FIT-S-23-8209 granted by Brno University of Technology.



