Comprehensive Study Of Malware Analysis on DNS over HTTPS(DoH)

Indian Institute of Technology Kanpur

Department Of Computer Science and Engineering

Project Report - Log4j
CS658A: Malware Analysis and Intrusion Detection
Under the guidance of
Prof. Sandeep Kumar Shukla
Academic Year 2022-2023

ABHISHEK SAHU (21111002)
(abhisheks21@iitk.ac.in)
Ashitosh Vankatrao More (21111017)
(ashitoshvm21@iitk.ac.in)
Binaya Kumar Suna (21111021)
(binayas21@iitk.ac.in)
Shivam Kharwar(21111058)
(skharwar21@iitk.ac.in)
Tanikella Sai Kiran(21111061)
(tskiran21@iitk.ac.in)
MAJ Ashish Ahluwalia (21111073)
(ashisha21@iitk.ac.in)

KEYWORDS: Domain Name System(DNS),DNS-Over-HTTPS(DoH),DoH Tunneling, Proxies, HTTPS, man-in-the-middle attacks (MitM),IP Spoofing, D DNS tunneled traffic characterization, encrypted traffic charcterization Malware analysis,DoH malicious packets,Static analysis

Chapter 1

Introduction

1.1 Motivation

Domain Name System (DNS) is the internet backbone for providing a mapping between human-readable hostnames and computer understandable Internet Protocol (IP) addresses. It uses simple UDP because of that, it contains some issues.

1. DNS tunneling is a method of using DNS protocol as a mean of encapsulating data transmission between a client and a server.

2. This vulnerability makes DNS protocol highly susceptible to various active and passive attacks, such as man-in-the-middle attacks (MitM) and eavesdropping.

In 2018, a new protocol named DoH was released, which not only enhances performance but also improves user privacy and security by preventing eavesdropping and manipulation of DNS data through MitM attacks. DoH wraps DNS records (both requests and responses) in an HTTPS connection, providing encryption and authentication of the server, and changing the connection-less aspect of DNS. DoH primarily serves two purposes–preventing on-path devices from interfering with DNS operations and allowing web applications to access DNS information via existing browser APIs.

1.2 Problem Statement

DoH has already been criticized by many security re-searchers for making DNS tunnels harder to detect and mitigate.

1. Since the DoH wraps the DNS traffic in HTTPS, the DNS traffic is imperceptible to the network infrastructure between the client (malware) and the DoH server. This effectively makes detection methods that rely on examining the DNS packets obsolete for the firewalls.

2. Since HTTP/2 is the minimum version of HTTP that DoH standard recommends for using with DoH, Malware can utilize the HTTP/2 connection to send several DoH request, without creating a separate connection (or packet) for each request.The same also applies to the responses that DoH server is sending to the malware.

3. Malware can hide the frequency of their DNS resolutions, further reducing the number of methods that can detect DNS tunnels

4. Our goal is to train a model which can detect DNS tunneling for DNS over HTTPS(DoH).

Chapter 2

Methodology

2.1 Data Collection

We have collected different PCAP file from CIRA-CIC-DoHBrw-2020 which contains HTTPS traffic flows with two levels of distinct labels i.e DoH and Non-DoH.

2.2 Data Preprocessing and Feature Extraction

We have done these following steps to

1. We have read the captured traffic in PCAP format which is created by tools such as tcpdump or Wireshark and extract the features.

2. The traffic captured in the dataset in form of PCAP files as input and extracts features for each flow in the input.

3. This process is done in image@cse.iitk.ac.in and it took roughly 20 days to extract the files.

4. Results are saved in a CSV file, where each row in the output CSV file would specify a flow in the input traffic.

5. Also we have done the labelling of the outputs CSV into Benign or Malicious based on the label of the PCAP files.

6. Then we have selected a subset of features based on our intuition for our training.

7. We have splitted our dataset into two parts training and testing in ratio 70:30.

2.3 Implementation

To acheive our goal, we have developed a two layer classification approach. In the first layer, we have developed a classification model which will divide our training data into DoH and Non-DoH packets. In the second layer, we characterized the DoH packets detected in layer 1 into Benign-DoH and Malicious-DoH. The two layers used for classification are explained below:

2.3.1 Layer 1

At layer 1 we will classify the packets in to DoH and non DoH using these following classifies

1. Random Forest (RF) : With parameters gini index, best splitter, min_samples_split =2
2. Decision Tree (DT) : With parameters gini index,best splitter,min_samples_split=2, max_depth=10, random_state=10
3. Gaussian Naive Bayes (NB) : With parameters var_smoothing = $1e - 9$
4. CNN Classifier : With parameters binary cross entropy loss, adam optimizer and Activation functions are Signmoid activation function and Rectifier linear unit, accuracy metric and kernal_size=3
5. DNN Classifier : With parameters binary cross entropy loss, adam optimizer and Activation functions are Sigmoid activation function and Rectifier linear unit, accuracy metric and kernal_size=3

2.3.2 Layer 2

After layer 1 classification, we have perform characterization of DoH packets. Here we will classify the DoH packets detected in level 1 them into Benign DoH and Malicious DoH.

1. Random Forest (RF) : With parameters gini index, best splitter, min_samples_split =2
2. Decision Tree (DT) : With parameters gini index,best splitter,min_samples_split=2, max_depth=10, random_state=10
3. Gaussian Naive Bayes (NB) : With parameters var_smoothing = $1e - 9$
4. CNN Classifier : With parameters binary cross entropy loss, adam optimizer and Activation functions are Signmoid activation function and Rectifier linear unit, accuracy metric and kernal_size=3
5. DNN Classifier : With parameters binary cross entropy loss, adam optimizer and Activation functions are Signmoid activation function and Rectifier linear unit, accuracy metric and kernal_size=3

2.4 Testing

2.4.1 Layer 1

We have tested our model on the testing dataset that we have generated for level 1 and the result produced is shown below.

Model	Precision	Recall	F-Score
Random Forest	0.993	0.993	0.993
Decision Tree	0.993	0.993	0.993
Naive Bayes	0.84	0.834	0.833
DNN	0.97	0.97	0.97
CNN	0.98	0.98	0.98

Table 2.1: Results for Level 1 Classification

2.4.2 Layer 2

After testing the model for layer 1 we tested for level 2 on same testing dataset and corresponding results are presented below.

Model	Precision	Recall	F-Score
Random Forest	0.999	0.999	0.999
Decision Tree	0.999	0.999	0.999
Naive Bayes	0.836	0.833	0.832
DNN	0.98	0.98	0.98
CNN	0.99	0.99	0.99

Table 2.2: Results for Level 1 Classification

Chapter 3

Conclusion

We have implemented a two-level binary classifier and then used the statistical function to detect DoH connections and Malicious DoH activities (DoH Tunneling). The First layer distinguishes DoH traffic from Non-DoH traffic. And the second layer determines malicious characterized and harmless DoH flows.

3.1 Limitations of our system

1. Some tools create the dataset that we have used to train our model, so it might not capture the same traffic flow generated by DoH tunneling traffic
2. For testing our model, we don't have any live data
3. There might be some other areas of DoH protocol that still inherit DNS vulnerabilities.
4. We have tested our model in static analysis

3.2 Improvements

1. Real-time dataset can be used to train our model
2. We can extend our model to dynamic analysis