Malware identification and classification using HTTP protocol

Malware (malicious software) is a serious threat to the modern Internet. Criminals use it to send unsolicited messages, extort ransom by encrypting files, or steal bank login credentials. Malware utilizes popular network protocols for communication purposes, including the frequently used HyperText Transfer Protocol (HTTP). The goal of this dissertation is to demonstrate that malware-generated HTTP requests can be used for the identification and classification of malware families. For the purpose of this study, network traffic datasets covering 121 malware families and a set of popular benign applications were created. The conducted experimental evaluation was divided into three parts. The first part identified characteristic features of HTTP requests which allow distinguishing between malware and benign applications. Then, these features became the basis for the second part of the analysis, which resulted in the creation of a tool called *Hfinger*, that can to create unique representations of HTTP requests. These representations can be used to identify malware by distinguishing its families as well as its specific operations, e.g., attacks or downloading commands. Finally, the third part of the study focuses on the problem of classifying malware using machine learning algorithms, i.e., assigning the names of specific families to the analyzed network traffic. The problem is extended by the recognition of classes that were unknown during the training phase of the classifier, which is also called Open Set Recognition problem. During the experimental evaluation, two HTTP request representations were used: the first one generated by *Hfinger*, and the second one using n-gram analysis. To the author's best knowledge, this is the first work focused on the application of the Open Set Recognition approach to classify malware based on its HTTP network traffic.

Keywords: malicious software, malware, network traffic analysis, HTTP protocol, classification, Open Set Recognition.