Information Security Institute

MSSI CAPSTONE PROJECT PROPOSAL

Malware Detection through Data Analytics

TEAM

Xiao Chong Chua (Cyrus Chua), Chen Cao

Advisors

Dr. Song Luo (JHUISI), Dr. David Silberberg (JHUAPL)

Introduction

In today's world, malware is a big problem for all sorts of computing devices, ranging from web servers to mobile phones. What exactly is malware? It is a combination of two words – "malicious" and "software". Software is considered malicious when it is created with the intent to be malicious, or if it results in activity that is detrimental to the system it is targeted at. (Aycock, 2006) In fact, according to McAfee's quarterly threat report released in March 2016, in the fourth quarter of 2015, there were almost 500 million types of malware in existence, and more than 40 million of them were new malware.

At the same time, data analytics is a field in computer science that is rapidly gaining traction, and being applied in a variety of ways to solve problems in businesses, science, cybersecurity, etc. In this capstone project we will seek to explore techniques and models commonly used in data analytics to detect malware present in network traffic.

The Problem

The main problem we have here is that many types of sophisticated malware often exhibit behavior that actively subverts detection on a network. Examples of such malware aim to bypass firewalls and intrusion detection systems without being detected, in order to gain entry to a protected network, infecting the computing devices within. Differentiating malicious traffic generated from malware from large amounts of legitimate traffic is almost like finding a needle in a haystack. However, because of rapid advancements in the field of data science, it is realistically possible to solve this problem by applying the right techniques.

Objectives

Since the timeframe of this project is only four months, we have to narrow the scope. The aim of the project would be to find a particular type or subgroup of malware, and to solve the problem of detecting this particular type or subgroup of malware by the use of algorithms and data analytics. The goal at the end of this project is to achieve a reliable method to filter out malware by detecting anomalies over the network. The main programming language we expect to use for this project is Python.

Milestones

	Description of task	Tentative date of completion
1	Literature review on malware behavior, ways to detect it, as well as applicable techniques used in data analytics.	9/25
2	Prepare or find a realistic data set to simulate an enterprise network infected with malware.	10/9
3	Mid-project progress report.	10/23
3	Design and implement detection algorithms.	11/6
4	Analyze the results and draw conclusions.	11/13
5	Write a technical report of our findings.	11/20

Citations

Aycock, J. D. (2006). Computer viruses and malware.

Threats Report (Tech.). (2016, March). Retrieved September 9, 2016, from McAfee website: http://www.mcafee.com/us/resources/reports/rp-quarterly-threats-mar-2016.pdf