CALIFORNIA STATE SCIENCE FAIR
2010 PROJECT SUMMARY

Name(s)

Ryan P. Batterman

Project Number

S1602

Project Title

Malware Identification by Statistical Opcode Analysis

Abstract

Objectives/Goals

This project determined the efficacy of statistical analysis of program assembly instruction (opcode) frequencies to identify Malware from Goodware.

Methods/Materials

Malware and Goodware binaries were obtained and a python script was created to extract opcode frequencies from specific parts of these files. Naive Bayes models and Kmeans based models were then trained using these executables. These models were tested using a different set of programs to determine their efficacy at identifying Malware from Goodware.

Results

The best Naive Bayes model had a recall of 1 for Malware and .8 for Goodware.

Conclusions/Discussion

Differences in opcode frequencies can differentiate Malware from Goodware. Certain instructions occur much more frequently in one group than in the other; these differences can be used to identify the two types of programs.

Summary Statement

This project examines models that differentiate Malware from Goodware using the frequencies of program assembly instructions.

Help Received

Communicated with mentor Joshua Kroll ; Pamela Durkee proofread papers and guidance