CERIAS

The Center for Education and Research in Information Assurance and Security

PURDUE UNIVERSITY

Understanding Malware through Classification and Machine Learning

Cory Q. Nguyen[$], Rui M. Esteves, Dr. J[&]. Eric Dietz[$], Dr. Thomas J. Hacker[**]

Purdue Homeland Security Institute[$]

Computer & Information Technology
High Performance Computing[**]

VeriSign
University of Stavanger[&]

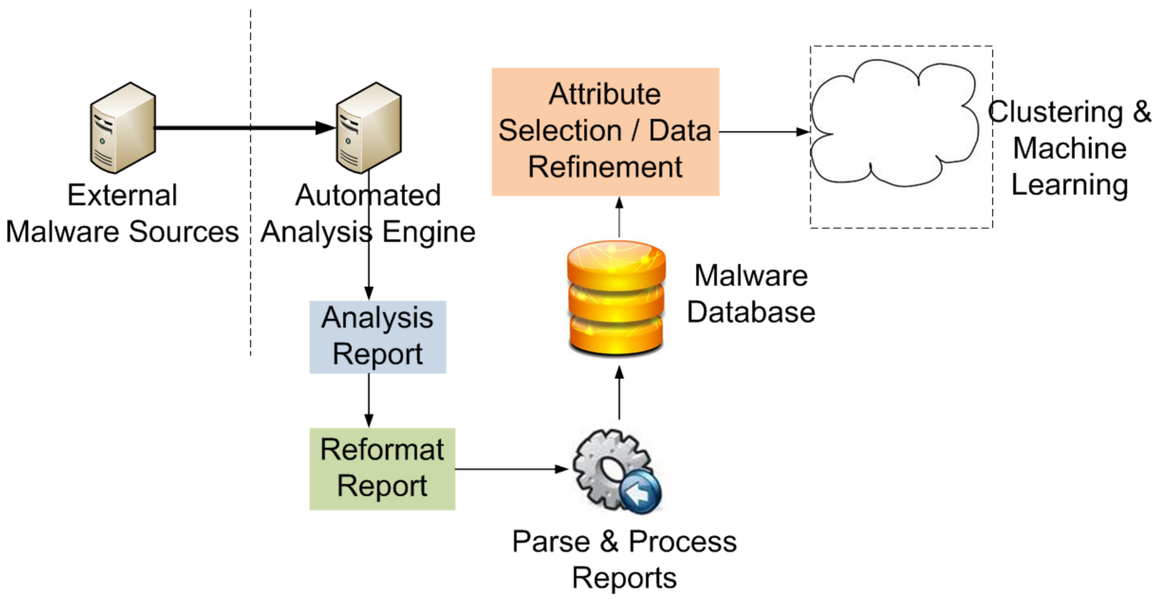
Overview

This study applies the concepts of machine learning and techniques of classification to help us better understand malware and its behavior and developing trends. In studying it's behavior and characteristic , not only do we have a better understanding of what an unknown malware does and what its purpose is, but also a better idea of how to respond to the threat. It is important to accurately classify and identify unknown malware to reduce the response time.

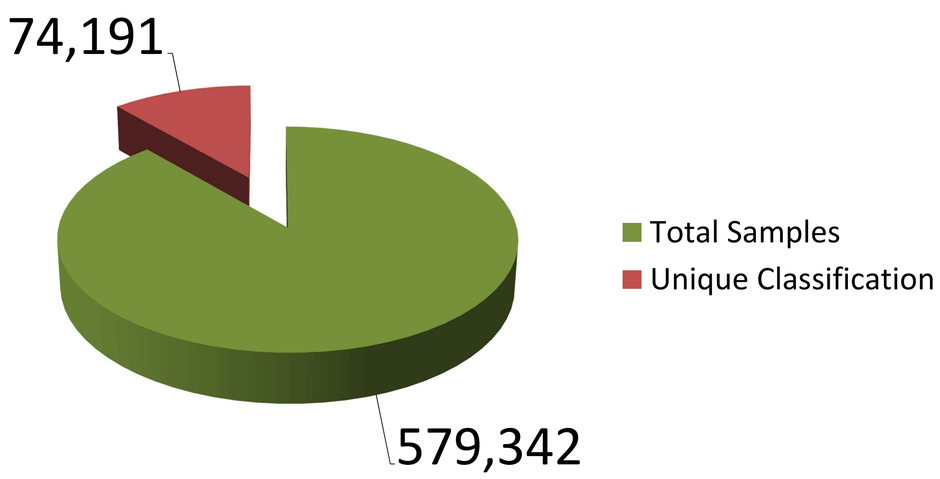
Currently, our lab has collected a total of 579,342 samples with 74,191 unique samples. These samples are processed through an automated analysis engine which outputs reports where are then refined, formatted, and indexed into a central database. Once the raw data is uploaded. Specific attributes are selected to be processed for learning and predictive modeling.

We have used clustering algorithms, specifically, a mixture of partitional and hierarchical clustering, however, our attribute list continue to grow by n-dimension. Thus, we are currently accessing how to shorten our attribute list. Previously our attribute list was composed of more than 59,000 attributes. Recently, we shortened the list to 23 attributes, however, the accuracy and feasibility of these 23 attributes have yet to be evaluated.

Current Research Process



Current DB of Malicious Binary



Different Categories & Number of Groups

Categories	# of Unique Classification
Adware	798
Worm	3067
Worm – P2P	500
Worm - IM	232
Worm - Email	87
Rootkit	2588
Trojan	54803
Trojan-Downloader	5520
Trojan-GameThief	974
Trojan-Clicker	283
Trojan-Proxy	130
Tools	
DoS	187
Flooder (Email, IM)	23