CSE 494: Artificial Intelligence for Cyber Security: Spring 2019

Lab 2: Malware analysis using VirusTotal reports on malware samples

Objectives of the lab:

- See where to extract MD5 hashes from and use them to access VirusTotal API – obtain reports on the malware samples.
- Learn about various attributes of malwares that would be used to characterize the samples based on their families or types
- Learn how to convert various attributes of the malwares into feature vectors suitable for data mining

The first part of the demo would show how to extract reports on malware samples using the ViruTotal API (20 min.) api key is in jupyter notebook

1. *Using the VirusTotal API* – To extract the malware reports based on MD5 hashes, we would be using the Private API of VirusTotal. You will be provided with the API Key that would be needed to access the reports based on the MD5 hashes.
    a. Sources from where MD5 hashes can be obtained based on OS, malware families or types
    b. Parse the JSON report obtained for the malware samples to extract the relevant information to run data mining algorithms.

using family to evaluate quality

The second part of the demo would show how to analyze malware samples using the VirusTotal API and clean the data for input to scikit learn (40 min.)

2. *MALWARE ANALYSIS* – (20 mins.)
3. You can donwload the list of malware md5 hashes list from the blackboard – *lab 2_md5_families.csv and lab_2_md5_types.csv*
    a. Analysis of malware sizes by families. The following 4 families will be used in this demo

        Dridex
        Locky
        TeslaCrypt
        Zeus

    b. Analysis of malwares by type. The following 4 types would be used in this demo

        Trojan
        Worm

Unsupervised Learning doesn't use families to compute the

clusters

virus

backdoor

Analysis would include the following distributions:

i. FileOS
ii. Codesizes
iii. Initialized data size
iv. Entry point
v. Subsytem

4. *Feature curation (20 mins.)*
   a. Categorize each of the attributes into categorical, ordinal, textual, binary attributes.
   b. Conceptual demo on feature vectorization, normalization, encoding, n-grams(optional in case of textual descriptions), n-grams for hex dump based features.
   c. Convert the features into a feature matrix that can be fed to various data mining algorithms – cases of missing data/dropping variables.