Project Proposal – Team 2

**Utilizing Machine Learning Techniques for Classifying Malware**

**Introduction**

The explosion of malware in recent years has created an enormous burden on cyber security professional. In 2014 more than 300 million new malware threats were created [1]. In an attempt of avoid detection by anti-virus software the authors of malware will often mutate (modify, obfuscate, etc.) their malicious software. This results in a single piece of malware looking like many different pieces of software (the behavior is the same but the source code looks different). Cyber security professionals need automated approaches to identify and group malware into families (similar behavior) to make it easier to combat the threats they pose.

**Overview**

A large malware (500GB+) data set was released by Microsoft in 2015 [2]. The data set consisted of more than 20,000 malware samples. Each malware sample has two formats; the raw hexadecimal representation of the binary and a metadata manifest (generated using the IDA disassembler tool). The malware samples are also categorized into 1 of 9 malware families.

This project will use machine learning techniques (Naïve Bayes, SVM, Decision Trees, etc.) to classify the malware samples into their respective families and evaluate which techniques perform the best. With the data set being so large we are anticipating the need to use Big Data solutions (AWS, Spark, Storm, etc.) to produce results in a reasonable amount of time.

**Approach**

Our project will consist of the following steps:

* Research related work
* Identify features to extract from the raw hexadecimal and metadata files
* Extract features from the data (use Big Data tools if necessary) to create training and testing samples
* Train models (Naïve Bayes, SVM, Decision Trees) using training samples
* Evaluate models using the testing samples

**References**

[1] <http://money.cnn.com/2015/04/14/technology/security/cyber-attack-hacks-security/>

[2] https://www.kaggle.com/c/malware-classification

[3]

[4]

[5]

[6]

[7]