Machine Learning
Project Plan

Shubham Tripathi
201407646

Title: Host-Based Intrusion Detection System (HIDS)

Problem Statement: Given normal and intrusive (attack) sequences of system calls, determine whether given test sequence is intrusive or normal.

Dataset: ADFA- linux dataset '14 [2]: Each trace represents a sequence of system calls. There are normal traces and attack traces. This is the latest dataset available publicly from 2014 for testing IDS systems.

THE COMPOSITION OF ADFA-LD

Normal	
# of Training traces	833
# of testing traces	4373
Total attacks	
# of attacks	60
# of attacks traces	686

Planned Methodology:

1. Applicability of Bag-of-words approach. Since the attacks are made as sequences of privileged commands the number and type of system calls used vary between normal and intrusive sets, this approach may be effective. Also the vocabulary(# system calls) is not large, data representation may not be a problem.

2. Study the approach devised by the security group who made ADFA dataset available in the paper: A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns.[1]

3. See why the IDS approaches that worked well on previous datasets like UNM and KDD98 fail to work for this ADFA dataset. I.e. how this dataset benchmarks the latest intrusion detection systems.

References:

[1] G. Creech and J. Hu. A Semantic Approach to Host-based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns. *Computers, IEEE Transactions on*, PP(99):11, 2013.

[2] G. Creech and J. Hu. Generation of a new IDS test dataset: Time to retire the KDD collection. In *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*, pages 44874492, 2013.
[3] G. Creech. Developing a high-accuracy cross platform Host-Based Intrusion Detection System capable of reliably detecting zero-day attacks, 2014