**Project Report for HMM: Anomaly Detection**

**(Homework 4: Question 3)**

a)Data sets used:

Normal dataset for training purpose (ps-live-unm)

Trojan dataset for testing purpose (ps-recovered)

b>HMM model:

A single HMM model is created for the multi time series given in the training set, i.e., same and one HMM model for all the processes.

Algorithms implemented

1>Forward

2>Backward

3>Baum Welch

c>The process involved in creating the HMM

Steps:

1>concatenate the two files (two training files into one file)

2>Get the ID’s of all the system calls and remove the duplicates to get the exact count of the system calls present.

3>Index each of the system calls (each system call just gets a single ID) stored.

4>Now create a hash map for all the actual sequence to the indexed system call, so when an observation is asked for, we get could get the direct access to the index pointing to the system call hence the state.

5>The number of states M = Number of observation, in this case 22.

6>Tackling the 3 rd problem of HMM : finding the right parameter set = (A, B, pi).

First create a random transition and emission matrix.

In a loop till we find the appropriate probability

Baum welch:

Forward process

Backward process

ComputeGamma

ComputeXI

loop

7>This creates the HMM model.

We then try to attack the problem 1 of HMM i.e., to find P(O|for each process, both, in the training and the test data.

8>To do this we call the forward algorithm on training data and test data, using he model we learnt using Baum Welch and get the likelihood (loglikelihood). The ouput would be similar to what is shown below

d>Where **list1.txt and list2.txt** are the training data and **test.txt** is the test data.

**list1.txt**

likelihood 150 is -115

likelihood 159 is -116

likelihood 999 is -303

likelihood 193 is -122

likelihood 182 is -120

likelihood 326 is -182

likelihood 3299 is -194

likelihood 395 is -110

likelihood 170 is -118

**list2.txt**

likelihood 150 is -115

likelihood 4120 is -183

likelihood 159 is -116

likelihood 4121 is -164

likelihood 387 is -178

likelihood 193 is -122

likelihood 182 is -120

likelihood 4190 is -164

likelihood 325 is -181

likelihood 3737 is -198

likelihood 738 is -162

likelihood 17832 is -103

likelihood 484 is -183

likelihood 4147 is -215

likelihood 170 is -118

**test.txt**

likelihood 150 is -176

likelihood 813 is -224

likelihood 780 is -350

likelihood 159 is -178

likelihood 336 is -258

likelihood 193 is -183

likelihood 798 is -204

likelihood 182 is -182

likelihood 799 is -367

likelihood 398 is -265

likelihood 170 is -180

From the above it can be seen that anything below -**150** could be called as ones that could be malicious