Description of Diploma or Master Thesis

**Computer Virus Recognition with Grammatical Inference**

**Problem Description**
Traditionally, *signature-based* detection mechanism is used to recognize computer viruses which basically matches the pattern of the byte sequences indicating malicious code. The signature repository is often referred as virus dictionary and the entries are made by malware analysts when they identify a malware sample. This process requires significant amount of human effort and it does not work in case of viruses with unknown signature. Therefore, machine learning mechanisms have well been applied in the field of virus detection by researchers due to their apparent potential in unknown virus recognition. In this thesis a particular branch of machine learning, *Grammatical Inference*, and its potential application in virus recognition will be investigated.

Grammatical Inference is an area of machine learning where formal grammar is learnt from set of example data to discover intelligible generalization of common properties of the given dataset. A canonical automaton is often created by means of various learning algorithms, which recognizes languages having similarity with the training set. Training a Deterministic Finite Automata (DFA) is often explored by researchers due to its position in Chomsky’s hierarchy.

**Scope of the Thesis**
In the first phase of the thesis, a state of the art analysis of the existing schools of thoughts of Grammatical Inference with respect to their suitability and adaptability in virus detection has to be conducted. Based on the outcome of the first phase one particular algorithm has to be implemented (possibly modified if required). Finally, experiments have to be conducted with real virus infected files.

The candidate is expected to have basic understanding of Automata Theory and profound English knowledge.

**Contact**
Hasan Ibne Akram
Room: 01.08.061, E-Mail: hasan@sec.in.tum.de, Tel.: (089) 289-18593
Start: Immediately