Extraction of features from binary files and creation of detection patterns

Ing. Marek Milkovič
xmilko01@stud.fit.vutbr.cz

In this work, we will deal with extraction of features from binary files and how to use them for creation of detection patterns while using the knowledge of theoretical computer science. Binary files form a very large group of files but scope of this work primarily focuses on executable files.

Malicious software is spreading around the world right now more than ever before. Therefore, it is necessary to study it in order to know how to detect and stop it before it can cause any damage either to user or his data. Detection patterns play a huge role in this scenario.

At first, we will show what kind of information we are interested in when dealing with executable files and how they can be extracted. For this task, we will present the usage of context-free language parser interconnected with semantic actions, but also parser based on scattered-context grammars [1].

As next, we will present what it takes to process extracted features and create full-fledged detection pattern out of them. Multiple approaches will be presented, like finding the longest-common subsequence, locality-sensitive hashing or creation of reduced finite-state automaton and further transformation to YARA language constructions.

Last but not least, the algorithms for matching of detection patterns will be presented. For that purpose, Aho-Corasick automaton will be introduced and how it can be used for pattern matching.

References

[1] J. Křoustek and D. Kolář. Context parsing (not only) of the object-file-format description language. 10:1673–1701, 10 2013.