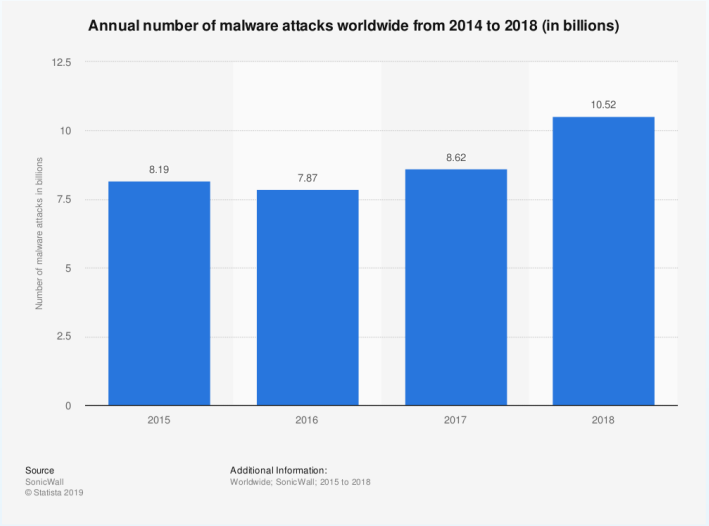
Malware detection algorithms based on machine learning and static malware analysis

Author: Bc. Lukáš Židzik Supervisor: Ing. Ján Hurtuk, PhD

Department of Computers and Informatics, Technical University of Košice, Slovakia

Motivation

Malware is a software with malicious behaviour. Nowadays, when variety of technical devices are all around us and information has a high price, it is a big security problem for whole society. Number of new malware samples is growing each year, so it is getting harder and harder to detect malware and defend against it's attacks. Used malware detection techniques work fine with known malware, but must be often updated because of new malware. And it takes a time. During this time malware is spreading and causing significant private and financial issues.



How to detect new malware?

We need to detect something we don't know how it looks or how it will behave. We only know that it will be similar to other malware in some features. Based on these features we have to predict if unknown sample is malicious or not. Human being can't analyze a huge amount of features and samples but computers can. It's the solution. Use machine learning for it.

Dataset

The first step in machine learning process is gathering data. Data are used for learning. Our trained model should classify unknown sample into one of two classes (malicious benign). We needed both classes for learning. We gathered 3064 benign and 3066 malicious samples.

Processing

Based on a huge number of samples we have to partially automatize data processing. We used Robotask tool which repeatedly do a defined tasks.

Feature extraction

There is a lot of features in each sample, but we have to focus on these which are specific for malware. Based on previous researches we choose 96 Win API functions, which malware use the most. We analysed each sample by Dependency Walker and found if selected functions were called or not.

Results

The best of our trained models was tested on 1226 samples, where 1223 samples were qualified correctly. It means model prediction reach 99,84% accuracy.

Conclusion

We trained a few model based on different algorithms. All of these models detect malware with high precision. This approach is completely safe and it doesn't need any database of known malware.