Malware Research in PDF Files

Shir Bentabou and Alexey Titov
Advisor: Ph.D. Amit Dvir and Ph.D. Ran Dubin
Ariel University,Department of Computer Sciences,40700 Ariel, Israel

Abstract

Cyber is a prefix used in a growing number of terms that describe new actions that are being made possible by the usage of computers and networks. The main terms in those are cyber crime, cyber attack and cyber warfare, all of those can be carried out by malwares.

Malware, or malicious software, is any software intentionally designed to invade, cause damage, or disable computers, mobile devices, server, client,or computer network. Malware does the damage after it is implanted or introduced in some way into a target's computer. Nowadays there are many distribution strategies for malwares and many programs are used as platforms. Some of these programs are in the user's everyday use, and seem pretty innocent. In our project we will focus on PDF files as a platform for malware distribution.

PDF, Portable Document Format, is used for over 20 years worldwide, and has become one of the leading standards for the dissemination of textual documents. A typical user uses this format due to its flexibility and functionality, but it also attracts hackers who exploit various types of vulnerabilities available in this format, causing PDF to be one of the leading vectors of malicious code distribution.

Users normally open PDF files because they have confidence in this format, and thus allow malwares to run due to vulnerabilities found in the readers. Therefore, many threat analysis platforms are trying to identify the main functions that characterize the identity and behavior of malicious PDF files by analyzing their contents, in order to learn how to automatically recognize old and new attacks.

The target of our work is to test and analyze how the use of machine learning methods can lead to effective recognition of malwares in PDF documents.