Visual Analysis of Arbitrary Binary Data

Spiker, Matthew (School: Jefferson High School)

This project is an attempt to solve the problem of often unknown data types when handling some sort of computer data. When handling computer data, there may often be files or streams that are corrupted, headerless, or otherwise unrecognizable format. By analyzing the data in a way ignorant of the underlying format, a visual “fingerprint” of sorts can be created from patterns in the stream, thus identifying its type. This is what the project attempts to do. A Python program reads a source stream, whether it be a file or network socket, and counts the frequency of consecutive byte patterns such as a 0xe6 followed by a 0x21. By identifying these patterns, the entire 256x256 array is plotted on a graph with color showing the occurrences of that byte pair, and this can be used to uniquely identify different file types. The results and distinctly showed that it is possible to use this visual analysis method to identify data formats. Raw data types showed up as bit streams in vivid patterns or grids; for example a .elf file shows a gridlike pattern associated with machine code. On the other hand, files with high entropy do not show detail; an encrypted file by definition does not show any patterns (which it didn’t). This field is called Visual Reverse Engineering and is a niche field within computer science and can have many applications from data recovery to malware analysis to everyday life by identifying an unknown file.