An approach for malware analysis using hashing

Malware is a short form of MALICIOUS SOFTWARE, which is a collective phrase for all software developed for disrupting, damaging or gaining access to data and systems in an unauthorized manner.

**Types of malware:** There are a number of different ways of categorizing malware; the first is by how the malicious software spreads. You've probably heard the words virus, Trojan and worm used interchangeably, these three can be described subtly in different ways malware can infect target computers:

1) A “worm” is a standalone piece of malicious software that reproduces itself and spreads from computer to computer.

2) A “virus”is a piece of computer code that inserts itself within the code of another standalone program, then forces that program to take malicious action and spread itself.

3) A “trojan”is a program that cannot reproduce itself but masquerades as something the user wants and tricks them into activating it so it can do its damage and spread.

Malware has remained a consistent threat since its emergence, growing into a plethora of types and in large numbers. Malware Authors have managed to increase their malware’s sophistication to avoid detection against anti-malware technics by implementing new features and specific modifications, such as encryption, polymorphism and metamorphism to maximize their resilience. Anti-virus vendors, researchers and analysts are struggling to keep up and to identify all the new malware types that are created every day. However, anti-virus vendors and analysts managed to adapt their identification techniques by relying on automated analysis methods, and tools in order to distinguish malicious from benign code, such as the traditional static analysis using Cryptographic hashes with the MD5 and SHA256 being the most commonly used in Malware Research in 2013.

Hash is a function that converts one value to another. Hashing is a practice of using an algorithm to map data of any size to a fixed length.

CRYPTOGRAPHIC HASHES: Cryptographic hashes have been the traditional tool, for both malware analysis and forensic investigations.

The most used Cryptographic Hashes for malware analysis to detect identical known malware signatures are the MD5, SHA-1, and SHA-256.

The main difference between these Cryptographic Hashes is their hash value length, despite they share the same concept to perform known file filtering by identifying identical matches. As mentioned earlier, changing a single bit in a file/document means that its hash value will also change, which makes it impossible- arguably- to find any associations or similarities between two files or to check if they are virtually similar.

1. Rolling Hashes: The Rolling Hashes are generating ‘pieces’ of the traditional hash strings by producing a pseudo random value based only on the context of the input. They are popular because they are easy and fast to compute. They are “used to identify similar strings in blocks of data.
2. Piecewise Hashing: The Piecewise Hashing generates a final checksum for the whole document like traditional hashes. They overcome the limitations and the drawbacks of the latter as piecewise hashes separates the whole file into fixed segments/pieces, then generates a hash values for each of these segments. The generated segments values are, at the end, forming the final hash sequence. Furthermore, they were initially created to reduce the potential errors during the forensic imaging, so that the integrity of the data will be absolute, because only one hash segment would be void.
3. Fuzzy Hashing: FUZZY HASHING Fuzzy hashing (FH), which is also called Context Triggered Piecewise Hashing (CTPH), is a combination of Cryptographic Hashes (CH), Rolling Hashes (RH) and Piecewise Hashes (PH). In fuzzy hashing analysis, the file of interest is divided into multiple blocks and a hash value is calculated for each block, with the final step being the concatenation of all hash values of the blocks to generate the fuzzy hash value. A number of factors affect the length of the fuzzy hash value, including the block size, the size of the file and the output size of the selected hash function. Fuzzy hashing methods can be classified into several categories: Context-Triggered Piecewise Hashing (CTPH), Statistically-Improbable Features (SIF), Block-Based Hashing (BBH) and Block-Based Rebuilding (BBR). Some of the methods of fuzzy hashing are SSDEEP, SDHASH, mvHASH-B and mrsh v2 Multi-resolution similarity hashing (mrsh)
4. Import Hashing: IMPHASH Import hashing is one of a number of methods used to ascertain the similarity of two files. This method utilizes import libraries (function calls from other software), where the order in which they are called and the functions themselves are utilized to generate a hash value.