B7 Fixed Size and Variable Size Block Hashes for File Identification

Douglas R. White, MS, 4225 Angell Road, Taneytown, MD 21787-2601*

After attending this presentation, attendees will understand some of the principles of identification of files during investigations of computer systems based on cryptographic hashes of files and partial files.

This presentation will impact the forensic community by introducing the rigor of cryptographic digital file identification at a granular level, which supports statistical identification of objects.

Use of cryptographic hashes or “digital fingerprints” to automatically identify files is absolute when applied to a file as a whole, where the file is unambiguously categorized. When dealing with morphing digital objects, such sorting leaves many files to be dealt with by manual review.

Block hashing is a method of applying the cryptographic algorithms to smaller-than-file size portions of the suspect data. Previous work used cryptographic hashes to “fingerprint” portions of data files, which assist investigators in identification of modified and partially deleted suspect data. Such cryptographic hashes cannot be used to identify similarities between data.

In this case study, cryptographic hashes and “spamsum” fingerprints of corresponding variable sized blocks are computed. The aggregation of the block hash values allow statistical probabilities of identification of suspect files, taking the dynamic nature of digital objects into consideration. The association of a cryptographic hash with spamsum combines positive file identification with a method of identifying similar file content. With this information, investigators can identify portions of uncataloged files which are similar to portions of known, cataloged files.

Examples of practical applications of this technique will be presented. Files from 90 computer systems within one organization were processed. Examples of installation of common software applications can be identified, despite installation modifications. Identification of shared documents can be identified, including edit changes. This work addresses the use of simple anti-forensics methods to defeat automated file identification.

Cryptographic Hash, File Identification, Block Hash

* *Presenting Author*