D34 Built for Speed: Using Bloom Filters for File Identification

Douglas White, MS, National Institute of Standards and Technology, 100 Bureau Drive Stop 8970, Gaithersburg, MD 20899-8970*

After attending this presentation, attendees will understand some principles of storing, accessing and sharing digital file identification data in Bloom filters.

This presentation will impact the forensic community by calling attention to the Bloom filter as a storage mechanism; a probabilistic algorithm to quickly test membership in a large set using multiple hash functions into a single array of bits.

A list of cryptographic hash values from NIST's NSRL RDS 2.13 was mapped 16-byte (MD5) and 20-byte (SHA-1) integers and concatenated them to form a binary file. This is the most compact form we have found that preserves order and allows perfect matching. The binary file can be used to determine if an MD5 or SHA-1 is known in the NSRL.

At this time, the Bloom filters with which we are experimenting are stored in 512MiB files. The files have a header, a 2^32 bit (512MiB) Bloom bit map section, and may contain data after the bit map.

Tools for manipulating the bitmap data will be discussed. The implementation varies in the number of Bloom vectors used - 16 vectors for MD5, 20 vectors for SHA-1. The effects of changing the Bloom key size, vector count and number of inputs will be explained.

There are benefits and pitfalls to both Binary tree and Bloom filter search methods, which will be covered in this discussion. Our math shows that a Bloom filter with 35-bit keys, using 20 vectors can store 1,000,000,000 SHA-1 values with a 1-in-100,000,000 false positive rate, and be stored in 4GiB. Other speed, storage and distribution benefits of Bloom filter use will be shown.

Bloom Filter, File Identification, Digital File Storage

** Presenting Author*