Foundstone
Professional Services A DIVISION OF McAFEE

Practical use of cryptographic hashes in forensic investigations

Pär Österberg Medina

06/26/2012

Agenda

- Cryptographic hashes
  - What they are and uses for hashes forensic
- Hash databases
  - How to look up a hash and create a database
- Hash collisions
  - Means to detect them
- Fuzzy hashing
  - How can we leverage “almost matching” in computer forensic

Who am I

- Pär Österberg Medina
  - McAfee and Foundstone Professional Service
  - Worked 8 years for the Swedish CERT
  - Background in Ethical Hacking

  - http://parosterbergmedina.blogspot.com/
  - https://github.com/parosterbergmedina

  - http://blog.opensecurityresearch.com/

What is a Hash?

- Wikipedia definition
  - A cryptographic hash function is a hash function, that is, an algorithm that takes an arbitrary block of data and returns a fixed-size bitstring, the (cryptographic) hash value, such that an (accidental or intentional) change to the data will (with very high probability) change the hash value. The data to be encoded is often called the "message," and the hash value is sometimes called the message digest or simply digest. http://en.wikipedia.org/wiki/Cryptographic_hash_function

Ideal cryptographic hash function

- Four important properties of a hash function:
  - Easy to generate the hash value
  - Impossible to know the hash in advance
  - Impossible to modify the data without changing the hash
  - Different data should not produce the same hash

Hash functions

- Cryptographic hash functions commonly used in Digital Forensic
  - MD5
    - Produces a 128-bit (16-byte) hash value
    - First published in 1992
  - SHA-1
    - Produces a 160-bit message digest
    - First published in 1995
  - SHA-2
    - A set of cryptographic hash functions (SHA-224, SHA-256, SHA-384, SHA-512)
    - First published in 2001

Hash in Forensic

- Used in Digital Forensic mainly for;
  - Authenticate evidence
    - Verifying the integrity of data
  - Identifying files by looking up the hash in a table
    - KnownGood
    - KnownBad
    - KnownUsed

Databases with hashes of files

- Downloadable databases with hashes
  - NIST - National Software Reference Library
    - Reference Data Set (RDS) with hashes for multiple OS and applications
  - Hashkeeper, AccessData, EnCase etc.

- Online resources
  - Bit9 FileAdvisor, SANS Hash Database, MHR from Team Cymru, Shadowserver Bin Check Service and many more

RDS format

- Content of a RDS archive
  - hashes.txt NSRLFile.txt NSRLMfg.txt NSRLOS.txt NSRLProd.txt

- RDS format

  "SHA-1","MD5","CRC32","FileName","FileSize","ProductCode","OpSystemCode","SpecialCode“
  http://www.nsrl.nist.gov/Documents/Data-Formats-of-the-NSRL-Reference-Data-Set-16.pdf

- Separating different ‘ProductCode’ values
  - Byte investigator - Tony Rodrigues http://sourceforge.net/projects/byteinvestigato

    # patch byteinvestigato/nsrlext.pl -i nsrlext-ParOM.patch -o nsrlext.pl

Looking up the hash

- Using ‘grep’ to search for the hash in a file
  - Old school and time consuming
- Indexing a file with ‘hfind’ from TSK
  - Still old school but less time consuming ;)
    - Supported types: nsrl-md5, nsrl-sha1, md5sum, hk

Build your own database

- Generating your own RDS compatible file
  - NIST Knopix distribution
    - Uses a Magic file for identifying the file type
    - Extract some archives; tar, gz, uu, iso, zip, cab, bz, bz2, rpm, inst and deb
  - Introducing hashdog
    - Relies on 7-Zip for extracting files
    - Generates RDS compatible files as well as other formats

How reliable are hashes?

- Hash collision
  - Two arbitrary inputs that will produce the same hash value
  - Reported in both MD5 and SHA-1
- Is then the SHA-1 and MD5 functions forensically sound?

Hashmap

- Introducing hashmap
  - Double mapping of hashes in a database
  - Indexed for greater speed

Fuzzy hashing

- Context Triggered Piecewise Hashing, also called Fuzzy Hashing
  - Introduced in SpamSum by Dr. Andrew Tridgell
  - Produces a Similarity Digests - Almost matching

Fuzzy hashing - implementations

- ssdeep - Jesse Kornblum
  - http://ssdeep.sourceforge.net/
- sdhash - Vassil Roussev
  - http://roussev.net/sdhash

Fuzzy hashing - uses in forensic

- Identifying files that are similar
  - fileA is looks a lot like fileB
    - Commonly used for identifying variants of Viruses
- Identifying clusters on the hard drive

Fuzzy hashing - hashdog

- hashdog will support the generation of;
  - ssdeep databases
  - sdhash database

This is the end

- Questions?