Hashing of File Blocks: When Exact Matches Are Not Useful

Douglas White

NIST United States Department of Commerce
National Institute of Standards and Technology

Disclaimer

Trade names and company products are mentioned in the text or identified. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Statement of Disclosure

This research was funded by the National Institute of Standards and Technology Office of Law Enforcement Standards, the Department of Justice National Institute of Justice, the Federal Bureau of Investigation and the National Archives and Records Administration.

Identification of Issues

- A meaningless change of file contents drastically changes a hash value
- Amount of data input to investigation is immense
- Common hash algorithms can not identify suspect files similar to known files
- Commonly used hash algorithms do not yield useful data on partial or deleted files

National Software Reference Library & Reference Data Set

The NSRL is conceptually three objects:

- A physical collection of software
- A database of meta-information
- A subset of the database,

  the Reference Data Set

The NSRL is designed to collect software from various sources and compute hashes of known applications. For the purpose of block hashes, we assume applications are benign.

Perturbing File Hashes

Use of cryptographic hashes to automatically identify files is absolute, too precise.

When dealing with morphing digital objects, such sorting leaves many files to be dealt with by manual review.

The NSRL hashset is commonly used to automatically remove benign known items from human processing, which is fail-safe.

Reducing Data Inflow

NSRL file content hash values allow investigators to automatically remove benign known items from view.

Known benign data can be identified before it arrives to investigators.

Is it technically possible to meaningfully reduce the amount of incoming data?

Block Hashes of Files

NSRL is investigating the usefulness of introducing the rigor of cryptographic digital file identification at a granular level which supports statistical identification of objects.

Block hashing applies the cryptographic algorithms to smaller-then-filesize portions of the suspect data.

File Selection

- NSRL investigated 4096-byte block hashes of Windows 2000 and Windows XP operating system files in our collection.
- NSRL also collected installed file block hashes from physical and virtual machines.

Block Selection

NSRL investigated 4096-byte block hash values.

4096 bytes was the smallest window considered, based on tools, storage and statistical applicability.

.2% of collection < 16KB

3% of collection < 32KB

27% of collection < 128KB

Block Hashing Benefits

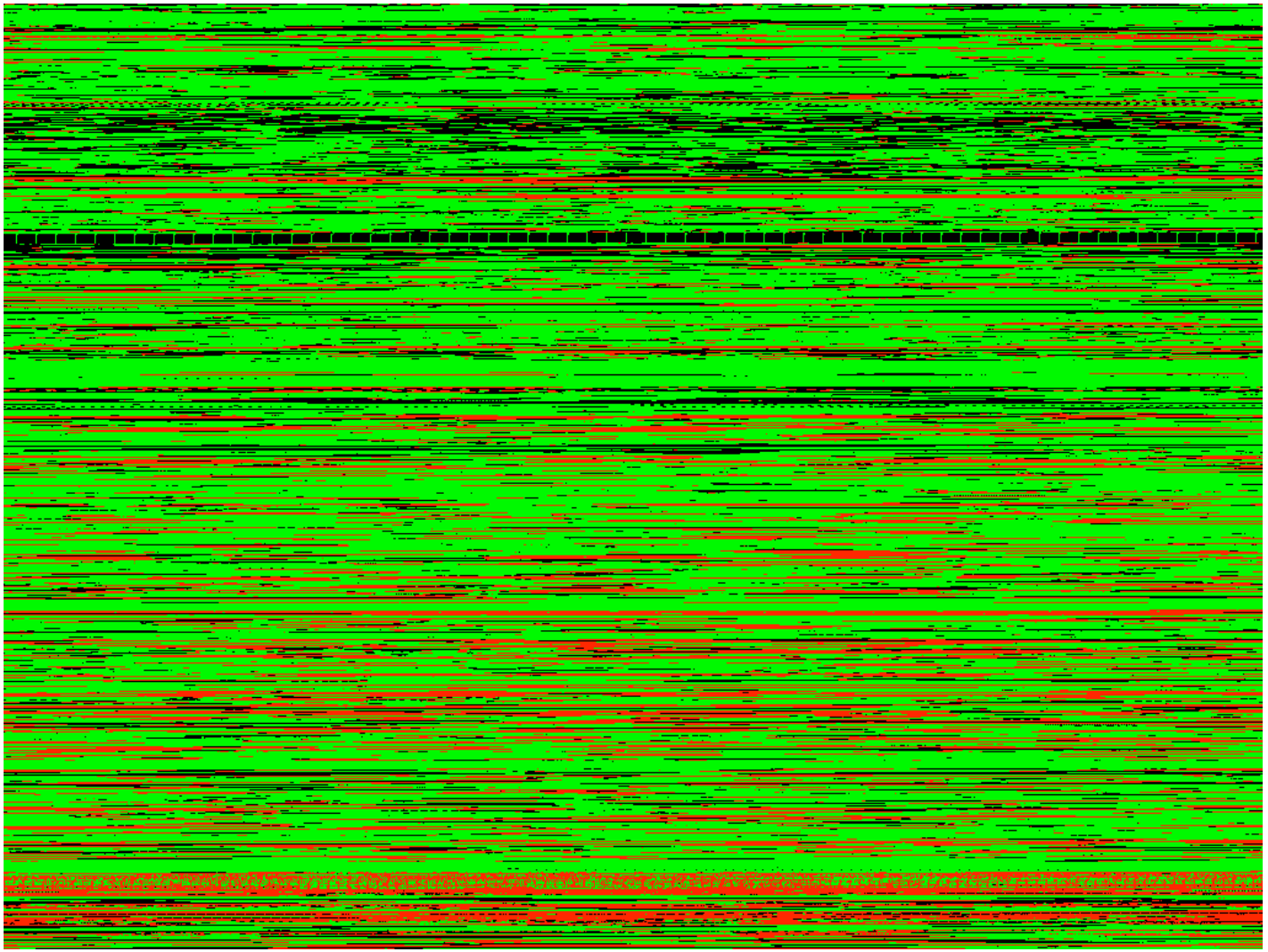
- File-based data reduction leaves an average of 30% of disk space for human investigation
- Incorporating block hashes reduces human review to 15% of disk space
- Assist in recognizing wiped media
- Assist in profiling media use

Physical and Virtual Machines

- P-XP vs. P-XP = 83% 8,679 files
- P-XP vs. V-XP = 85%
- V-XP vs. V-XP = 91%
- P-W2K vs. P-W2K = 85% 7,688 files
- P-W2K vs. V-W2K = 89%
- V-W2K vs. V-W2K = 94%

Known - Unknown - Zero

2nd 512 MB in W2K NTFS VM



Next Steps

Investigate a wider variety of applications

Automation & virtualization of installation

Comparison with “fuzzy” hashes

Storage in Bloom filter

Prototype disk block imager

“Smart unpacking” of remaining data

Contacts

Douglas White
www.nsrl.nist.gov
nsrl@nist.gov

Barbara Guttman
Software Diagnostics & Conformance Testing Division
barbara.guttman@nist.gov

Sue Ballou, Office of Law Enforcement Standards
Rep. For State/Local Law Enforcement
susan.ballou@nist.gov