NIST's
National Software Reference Library

Douglas White,

Software Diagnostics &
Conformance Testing Division

NIST
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

www.nsrl.nist.gov

National
Software
Reference
Library

Agenda

- Project background
- Mathematical hashes
- Law enforcement application
- NARA Presidential research
- Ongoing research

Project Background

- Collection is software application files
- Metadata database covers 45M files
- Metadata is publicly available
- Database schema, field descriptions

Project Background

- Files recursively harvested from media
- 5.25 FD, 3.5 FD, CD, DVD
- 4TB of data
- 12GB of metadata
- Mathematical hashes are the focus
  - SHA-1, MD5, MD4, CRC32

Mathematical Hashes

- Like a person's fingerprint
- Uniquely identifies file based on contents
- Primary hash value used is Secure Hash Algorithm (SHA-1) specified in FIPS 180, a 160-bit algorithm
  - 10^{45} combinations

Mathematical Hashes

- “Computationally infeasible” to find two different files less than 2^{64} bits in size producing same SHA-1
  - 2^{64} bits is one million Terabytes
- August 2004, CRYPTO conference
  - MD5 collision : specific initial values
  - SHA-0 collision : 80,000 CPU hours
  - SHA-1 collision : through 50 of 80 rounds
  - NOT pre-image attacks
- SHA-256 supercedes SHA-1 in 2010

Law Enforcement Application

- Automated elimination of benign application files from investigation
- Positive identification of “interesting” files
- Forensic tools use various metadata
- NIST provides unbiased court-admissible data to NIJ, FBI, DoD

NARA Presidential Research

- 93 subject computers
- 51,146 files totalling 2.3GB
- 11,118 unique files, 78% duplicate
- 8,077 files identified by SHA-1 (72%)

NARA Presidential Research

- 469 identical temporary installation files
- 161 zero-byte empty files
- 130 identical WordPerfect icon files
- Twenty other files have 90+ instances

NARA Presidential Research

- Possible to generate a “baseline” computer system
- Possible to obtain pedigree of operating system upgrades
- Possible to apply installed application metadata for further identification

Ongoing Research

- On-line archive of data
  - 2004 = 4TB, 2007 = 50TB
  - Apply new algorithms to collection
- Identification of file types
  - Over 3,500 identifiable
  - Reference data set of content files

Ongoing Research

- Metadata collection
  - Higher level of automation
  - Finer grain than file objects
  - Ad-hoc cluster of commodity hardware
- User access

Further Information

- Visit www.nsrl.nist.gov
- Email nsrl@nist.gov