NIST National Software Reference Library (NSRL)

Douglas White

September 28, 2005

Dwhite@nist.gov www.nsrl.nist.gov

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.

Introduction

The NSRL project is supported by the U.S. Department of Justice's National Institute of Justice, federal, state, and local law enforcement, and NIST. Other federal agencies and industry organizations provide resources.

The NSRL is designed to collect software from various sources and incorporate file profiles computed from this software into a Reference Data Set (RDS) of information.

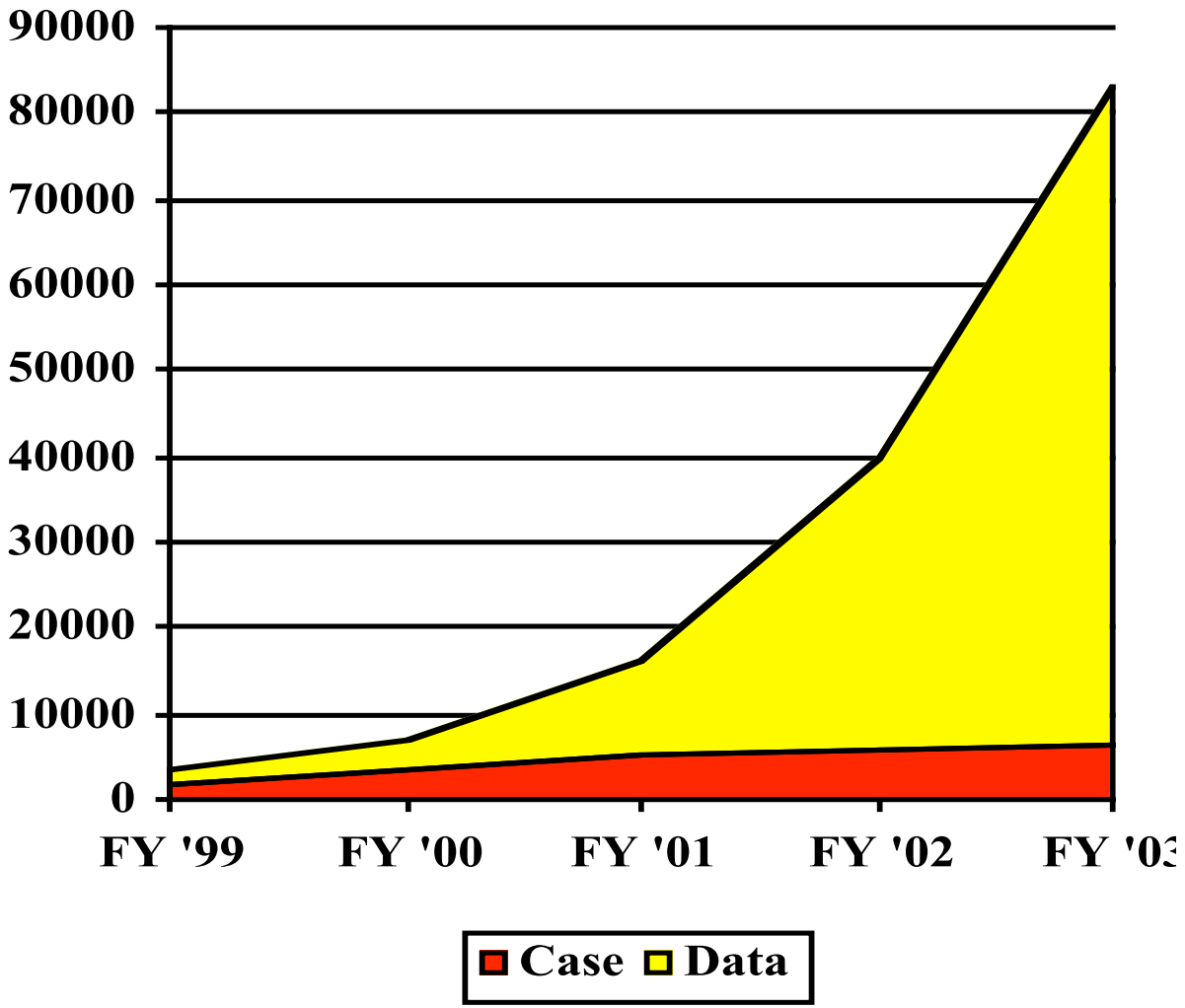
Computer Forensics Partners

- NIST, Office of Law Enforcement Standards
- DoJ, Nat'l Institute of Justice, FBI
- DoD, DCCC
- DHS, ICE, USSS
- State & Local Law Enforcement
- Vendors

Data Set Contents

The RDS is a collection of digital signatures of known, traceable software applications. There are application hash values in the hash set which may be considered malicious, such as steganography tools. There are no hash values of illicit data.

RDS 2.10 was available September 2005, providing 10,663,650 unique SHA-1, MD5 and CRC32 values for 33,860,009 files.



FBI's Cyber Caseload and Dataset Size Growth

Source: FBI CART, Oct 2003

Basics

The NSRL is conceptually three objects:

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

Physical Collection

The collection is treated as case evidence.

Software is kept in a locked room with limited access.

If metadata is questioned in court, it can be regenerated from original media.

The collection includes 7,009 applications, in over 35 languages, for many OSes.

Based on “popular” titles – most encountered, most pirated, most applicable at the time.

Non-NIST Physical Collections

There are collections off-site that NIST could specify as “verifiable” - National Archives, Library of Congress, National Library of France, Dutch Forensic Institute, IRS.

NIST can be a clearinghouse for unverifiable “hashes of interest” based on files that cannot be traced to original media, e.g. website downloads, one-off CDs.

Investigators can choose level of rigor needed - court admissible, peer reviewed, etc.

Downloads - legal issue for RDS, technically possible to collect.

Database

The database contains over 50 million file signatures.

All of the metadata is stored to uniquely identify a file in a directory on a piece of media in an application.

The hashes are only from files on original media.

Archive-type files (CAB, ZIP, UU, TAR, ISO, DMG, dd) are hashed, then extracted and contents are hashed.

Hash Collision News

- **The NSRL project does not see any fatal ramifications from the collision announcements**.
- Details posted at http://www.nsrl.nist.gov/collision.html
- We have not seen a "pre-image" attack; that is, the researchers did not identify a known file in the NSRL and attempt to generate a different file with a matching hash value.
- Nothing presented at Crypto 2004/2005 indicated that SHA-1 has been broken
- There are known MD5 collisions and weaknesses; the NSRL data provides an MD5 to SHA-1 mapping to facilitate the migration away from MD5.
- SHA-1 will be superceded in 2010 by FIPS 180-2, Secure Hash Standard (SHA-224, 256, 384,512). The NSRL will provide a SHA-1 to SHA-256 mapping.
- The NSRL provides several hash values and the file size, and it is highly improbable that a pre-image attack will be found soon that can generate a combination of hash collisions.

Database

Currently Windows system, running SQL server, using 12 GB for tables. Table dumps are available.

Evaluating DB schema, available on website; seeking input for data elements.

Database Redesign

Current database does not handle many-many relationships well; cannot respond to complicated queries spanning category, language, OS.

Plan rigorous taxonomies for category, OS, platform data.

Need to collect machine-based media identifiers.

Heading toward schema that lends well to automated (web-driven?) query building.

Reference Data Set

Version 2.10 was “released” to NIST duplication service September 6.

4 CD set contains zipped flat text files according to public spec on website. Can be imported into popular forensic tools.

Contains 10,663,650 unique SHA-1/MD5 values.

Expect 250K-1M new unique values per quarterly release.

NSRL Impact

- Referenced in 2001 seizure of bogus MS media in CA.
- Referenced by Simpson Garfinkel in 2002 efforts with reclaimed disks.
- Imported into EnCase, FTK, Ilook, Hashkeeper, Maresware.
- Essential to FBI CART, copied for every field office.
- Used by private organizations to eradicate P2P use.
- Used by ISPs to track app sharing on servers.
- Used by sysadmins to confirm valid OS file state.
- Used by FDA in FL Botox case.
- International use - UK NHTCU, EU JRC, etc.

NDIC Hashkeeper

- DoJ's National Drug Intelligence Center (NDIC) HashKeeper project produces hashsets
- Based on seized data and original media
- http://groups.yahoo.com/group/hashkeeper
- Three main FTP sites
- Over 300 hash sets

Other Hashset Sources

- Maresware
- Tripwire FSDB
- Hashkeeper, CFTT, iLook, CFID email lists
- Professional connections

Original Metadata Intent

The project sponsors were initially concerned with identification of known application files, to allow known files to be ignored, focusing investigation on user-generated data.

NIST does not assign “malicious” nor “notable” values to applications.

Evolving Metadata Intent

The NSRL does assign application categories, e.g. image manipulation, steganography, encryption. Original directory/path location is noted.

The NSRL metadata has been used to determine the “pedigree” of NARA systems. Can determine the upgrade path of a PC such as from NT3.5 to NT4 to W2K.

Other requested data are original MAC date/time, alternate data streams, byte signature info (Unix “magic”)

NSRL Research

By 2010, SHA-256 will supercede SHA-1 for Federal use.

NSRL will be collecting SHA-256 hashes from all media. Probably also Whirlpool and SHA-512.

Since we need to physically touch the media, we will be storing forensic copies on RAID. Current shelf contents are approximately 3TB.

Hot storage will allow NIST to perform future functions automatically on complete collection. The RAID will not be publicly accessible - copyright & NDA issues

NSRL Research

Have performed minimal installation hashes. None are in the RDS yet. Have captured installs in dd images and in virtual machines.

Researching block hashes - application of crypto hash strength to less-than-file granularity for statistical identification. Collecting hashes of 512B blocks. Applicable to dynamic files, slack space and deleted files.

Given amount of storage needed for block hashes, have investigated Bloom filters. This has certain advantages over binary tree storage.

Binary Search

Given collection of knowns, test against midpoint, if no match, test against midpoint of lower/upper half, etc. until match or can't recurse.

Can store 10 million MD5s in 160MB.

Takes 22 tests (10^7 items ~ 2^{22} items) for unknown.

Average of 11 tests for known MD5.

Bloom Filters

Essentially a bit vector, with three variables - key size, number of keys, number of items inserted.

Take MD5 value d41d8cd98f00b204e9800998ecf8427e

Convert the 8-hex-char chunks to 32 bit integers, set the bit at that position.

32-bit key forces 2^{32} bit vector (512MB), 100M items.

1 test if unknown, 4 tests if known (or false pos).

NSRL Research

Potential to store knowledge of 1 billion MD5 values in a 4GB Bloom filter, with 1-in-100-million false positive rate.

Block hashes and Bloom filter relevant to “far upstream” data reduction, possibly during disk imaging.

In process of migrating entire system core to open source products.

Agency with OS X/Linux/Apache/PostgreSQL/Perl/PHP talent could duplicate NIST NSRL results.

NSRL Research

Planning to make resources publicly available on the net; web queries, ODBC access, bulk custom subsets.

Developed a Knoppix-based boot CD distro to allow ad hoc cluster building.

NSRL cluster nodes can do arbitrary processing, beyond hashing - steg prediction, decryption, FFT image comparisons, etc. Image, audio content is useful despite content change, hash is of minimal value - other algorithms?

Collection Bottleneck

French software publishers must provide the National Library of France with copies of applications - there is no equivalent law or collection point in the U.S.

The Election Assistance Commission worked with voting software vendors to populate the NSRL with that category.

NSRL receives every MSDN media, Apple Developer media, and budgets $2,500 per month for purchases - a drop in the bucket. Have canvassed for donations in past with mixed results.

Hash Processing Capability

NSRL runs on dedicated, isolated 100Mbit network.

Have 1Gb hubs, NICs in critical locations.

Windows shares limit us to 12 drives for reading media.

Current setup can process 15GB per hour, media to hashset.

Will use fiberchannel in new rack-based system.

Move to Linux/OS X Samba shares allows more read drives.

New hashing nodes will be 64-bit dual CPU blades, should quadruple throughput out of the box to 60GB/hr.

Easy to add input drives, hashing blades for growth.

Contacts

Doug White
www.nsrl.nist.gov
nsrl@nist.gov

Barbara Guttman
barbara.guttman@nist.gov

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