2003 Cyber Crime Conference

Effectiveness of Hash Sets

Douglas White

National Institute of Standards & Technology

Overview

- Computer Forensics at NIST
- Available Hashsets
- Efficiency Questions
- Research Findings
- Future Research

Computer Forensics Partners

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

Role of NIST

- Mission: Assist federal, state & local agencies
- NIST is a neutral organization – not law enforcement or vendor
- NIST provides an open, rigorous process
- NIST data is traceable, court-admissible

NIST NSRL

- National Software Reference Library (NSRL)
  - Physical library of software, 2,603 products
  - SQL Server database of known file signatures
  - Reference Data Set (RDS): 17,900,000 file signatures
- Goals
  - Automate the process of identifying known files on computers used in crimes
  - Allow investigators to concentrate on files that could contain evidence (unknown and suspect files)

NSRL Reference Data Set (RDS)

- Reference set of file profiles
  - Each profile includes file name, file size, 3 file signatures (SHA1, MD5, CRC32), application name, operating system, etc.
  - Extracted from files on original software CDs, diskettes, and network downloads
- "Known" files – **not "known good"**
- 4,300 separate hashsets on www.nsrl.nist.gov
- Import into Encase, Ilook, TASK, 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

Questions on Hashset Efficiency

- How well do media file hashes identify installed files?
- How well do installed file hashes identify installed files?
- How does a physical installation or a virtual installation affect a hashset?
- How does rebooting a machine affect an installed file hashset?
- How well can hashset technology be expected to assist in an investigation?

Media File Hashset

- Win2000 Pro CD
  - 16,658 files, 16,539 unique SHAs
- Win 98 SE CD
  - 17,812 files, 17,436 unique SHAs
- Win XP Pro CD
  - 19,530 files, 19,404 unique SHAs
- MS Office 2000 CDs
  - 25,245 files, 12,495 unique SHAs
- Corel WP Office 2000 CD
  - 21,602 files, 21,554 unique SHAs

OS Virtual Installation

- W2K virtual install : 6,197 files
  - 4,625 unique files in installation
  - 3,679 (79%) identified by RDS
- W98 virtual install : 4,547 files
  - 3,937 unique files in installation
  - 2,993 (76%) identified by RDS
- WXP virtual install : 9,128 files
  - 6,547 unique files in installation
  - 0 identified by RDS
- Save snapshots on CD/DVD for repeatability

OS Physical Installation

- Physical machine install of W2K: 5,982 files
  - 4,340 unique files in installation
  - 2,621 (60%) identified by RDS
- Physical machine install of W2K: 5,982 files
  - 4,340 unique files in installation
  - 2,712 (62%) identified by VM install hash
- Virtual machine install of W2K: 6,197 files
  - 4,625 unique files in installation
  - 2,700 (58%) identified by PM install hash
- Save dd images on CD/DVD for repeatability

Application Virtual Installation

- MS Office 2000 on W2K
  - 18,903 unique application files installed
  - 18,682 found by RDS (98%)
- Corel WP Office 2000 on W2K
  - 2,846 unique application files installed
  - 2,356 found by RDS (82%)

Reboot Perturbation

- Virgin W98 system
  - Identified 13 files perturbed by reboot
- Virgin W2K system
  - Identified 6 files perturbed by reboot
- Virgin WXP system
  - Identified 16 files perturbed by reboot

Real-World Application

- Windows 2000 Pro
  - RDS hashset identified average 63% OS files
  - VM hashset identified average 61% OS files
- Windows XP Pro
  - RDS hashset identified 0 OS files
  - VM hashset identified average 60% OS files

HashKeeper

- Used hashset Z00166 for W2K Pro
- 998 unique MD5 values identified
- Identified 198 files (4%) on VM
- Identified 193 files (4%) on PM
- Both cases above identified 60+ files not identified by RDS

Physical vs. Virtual Machines

- Differences appear to be due to devices
- Virtual machines use abstract/generic device interfaces
- Physical machines require vendor specific drivers
- Need more data from installs on various hardware

Virtual Machine Software

- Most testing so far used VMWare 4
- Some testing used MS Virtual PC
- Moving to Bochs
  - Open source, free package
  - Can distribute configuration files
  - Better snapshot capability
- Raw data, procedures on www.nsrl.nist.gov

Technology Limits

- Eliminate known files on seized machine
  - Ranges from 60% to 75% for OSes
  - Ranges from 80% to 95% for applications
  - Intended use of RDS by LE
- Identify known files on seized machine
  - Only as good as the hashed collection
  - Intended use of HashKeeper by LE
- Dynamic files - path/name, byte sig, block hash
- Audio, images easily changed

Further Research

- Combinations of service packs, patches
- Automated install/restore process for VMs
- Automated install/restore process for PMs
- Block size hashing
- 3rd party tool API during hashing

Contacts

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

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