November 2001

ITL Bulletin

ADVISING USERS ON INFORMATION TECHNOLOGY

COMPUTER FORENSICS GUIDANCE

By Gary E. Fisher, Software Diagnostics and Conformance Testing Division, Information Technology Laboratory, National Institute of Standards and Technology

Introduction

NIST's Information Technology Laboratory has two ongoing research projects in computer forensics: the National Software Reference Library (NSRL) project and the Computer Forensics Tool Testing (CFTT) project. Funded through NIST's Office of Law Enforcement Standards (OLES) by the National Institute of Justice (NIJ), the Federal Bureau of Investigation (FBI), the Department of Defense Computer Forensics Laboratory (DCFL), the Department of Justice's Technical Support Working Group (TSWG), and NIST, these projects are designed to provide research into the efficient and effective use of information technology applied to computer forensics.

The NSRL provides a set of reference data that can be used to reduce the number of files that have to be reviewed or examined during an investigation. This can significantly cut down on the amount of time required to gather and verify evidence in crimes involving computers.

The objective of the CFTT is to provide a measure of assurance that the tools used in computer forensics investigations produce accurate results. Since the computer forensics arena is relatively new in terms of computer history, there are few standards available to define how these tools should operate and perform. The task of the CFTT is to develop specifications and test methods for these tools. The results provide the information necessary for users to make informed choices about acquiring and using computer forensics tools.

This *ITL Bulletin* presents some of the results of the work on these two projects and provides recommendations on how the results can be applied.

The National Software Reference Library Project

A single desktop computer can contain 25,000 to 100,000 stored files. A large percentage of these files are program executable files, library files, compressed files used to install applications, operating system files, etc. On a specific computer, these files can make up from 25 percent to 95 percent of the total number of files. Reviewing these files for evidence can take hundreds of staff-hours. In most cases, these files do not contain evidence, but without some automated process to assist in the review process, an investigator must review each file and make that determination.

The objective of the NSRL is to provide a foundation for automating the examination process by separating these files into those that are relevant to the investigation and those that are not. This is accomplished by computing a unique identifier for each file based on the file's contents. These identifiers can be used as file signatures or "fingerprints" for the associated files. The identifiers can then be compared to entries in a database of known fingerprints. If a file's fingerprint matches one in the database, it is a known file and can be eliminated automatically from examination. If the fingerprint does not match anything in the database, then the file is unknown and should be examined further. How this process is accomplished is described below.

NIST has been collecting software from various sources over the past 18 months. This software is recorded as the original source for known files and stored as a permanent part of the NSRL. The software in each package is stored on CD, diskette, magnetic tape, or other electronic storage medium. For example, one CD may contain a full implementation of an operating system. Another may contain a newer version of the same operating system. Other packages contain database

Continued on page 2

ITL Bulletins are published by the Information Technology Laboratory (ITL) of the National Institute of Standards and Technology (NIST). Each bulletin presents an in-depth discussion of a single topic of significant interest to the information systems community. **Bulletins are issued on an as-needed basis** and are available from ITL Publications, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8901, Gaithersburg, MD 20899-8901, telephone (301) 975-2832. To be placed on a mailing list to receive future bulletins, send your name, organization, and business address to this office. You will be placed on this mailing list only.

Bulletins issued since June 2000

- □ *Mitigating Emerging Hacker Threats*, June 2000
- □ *Identifying Critical Patches with ICAT*, July 2000
- □ *Security for Private Branch Exchange Systems*, August 2000
- □ *XML Technologies*, September 2000
- □ *An Overview of the Common Criteria Evaluation and Validation Scheme*, October 2000
- □ *A Statistical Test Suite for Random and Pseudorandom Number Generators For Cryptographic Applications*, December 2000
- □ *What Is This Thing Called Conformance?* January 2001
- □ *An Introduction to IPsec (Internet Protocol Security),* March 2001
- □ *Biometrics– Technologies For Highly Secure Personal Authentication*, May 2001
- □ *Engineering Principles for Information Technology Security*, June 2001
- □ *A Comparison of The Security Requirements for Cryptographic Modules In FIPS 140-1 AND FIPS 140-2*, July 2001
- □ *Security Self-assessment Guide For Information Technology Systems*, September 2001

management software, photo editors, word processors, image libraries, network browsers, compilers, accounting packages, and many other types of software. The concept is to collect as many different examples, versions, and updates of software as possible in order to generate file signatures for as many known files as possible.

Each file within a package, which may contain many thousands of individual files, is “fingerprinted” by passing the file through a program that computes a hash code. A hash code is a large number computed from the entire string of bits that form the file. The hash code is computed in such a way that if one bit in the file is changed, a completely different hash code is produced. To minimize the possibility that two different files may generate the same hash code, a sufficiently large hash value is computed.

The primary hash value used in the NSRL Reference Data Set (RDS) is the Secure Hash Algorithm (SHA-1) specified in Federal Information Processing Standard (FIPS) 180-1. SHA-1 is a 160-bit hashing algorithm, meaning that the hash value derived from the algorithm is 160 bits in length. The probability that two different files will produce the same SHA-1 value is 1 in 10^{80}, a very small probability. Several other standard hash values also are computed for each file. These include Message Digest 4 (MD4), Message Digest 5 (MD5), and a 32-bit Cyclical Redundancy Checksum (CRC32). These allow the SHA-1 values to be cross-referenced by other products that depend on different hash values. Additionally, this further ensures that no two files will have the same set of hash values. The hash values, file name, file size, and information identifying the source of the file are stored in the RDS. A separate, parallel, and independent process is used to validate the results of the primary RDS implementation. This ensures that the hashes computed can be verified to identify specific files in the RDS. Once verified and validated, the RDS is written to a master CD, duplicated, and distributed through NIST’s Standard Reference Data Office as Special Database #28 (http://www.nist.gov/srd/nistsd28.htm).

Who we are
The Information Technology Laboratory (ITL) is a major research component of the National Institute of Standards and Technology (NIST) of the Technology Administration, U.S. Department of Commerce. We develop tests and measurement methods, reference data, proof-of-concept implementations, and technical analyses that help to advance the development and use of new information technology. We seek to overcome barriers to the efficient use of information technology, and to make systems more interoperable, easily usable, scalable, and secure than they are today. Our web site is http://www.itl.nist.gov/.

When a computer hard disk, CD, or other storage medium becomes part of an investigation, the files stored on it can be “fingerprinted” using SHA-1, MD4, or MD5 through separate software that may be acquired from other sources. These fingerprints can be compared to the known file fingerprints in the RDS. Those files that have matching hash values can be discarded from the investigation without further examination; those that do not match the database should be examined further.

Uses of the Reference Data Set

There are several ways in which the RDS can be used depending on the goal of the investigator. Foremost is the highlighting of unknown files on a subject computer. Appropriate software can provide a list of files from the subject computer with file signatures that do not match any signatures in the RDS. An investigator can then review each of these unknown files for evidence while discarding the known files without having to review them. In test runs, the RDS has identified between 40 percent and 95 percent of the files on subject computers as known files, thus reducing the amount of time required for review.

Organizations are looking at the RDS for other purposes. For example, the RDS can be used in intellectual property investigations to find pirated software on a specific computer. If the pirated software is on the computer, i.e., a specific file or set of files exists on the computer, hash values of those files will match known files in the RDS.

Alternately, a perpetrator may try to hide a pornographic image by renaming it as a nondescript operating system file, e.g., renaming a .JPG image as an .EXE file. The hash value derived from the image will not match that from the known operating system file and will thus be uncovered.

Additionally, the RDS can be used to find camouflaged software, such as executable files with changed names that may blend in with other data files, e.g., a .EXE file renamed with a .JPG extension. A hash value of the camouflaged file will still match the RDS since the file’s contents were not changed, only its name.

Expected files may be missing if they do not show up in the known files list. This may indicate that files were deleted to cover up illegal activity and may prompt the investigator to pursue other means of investigating the file system.

The project website is http://www.nsrl.nist.gov.

ITL Bulletins Via E-Mail

We now offer the option of delivering your ITL Bulletins in ASCII format directly to your e-mail address. To subscribe to this service, send an e-mail message from your business e-mail account to listproc@nist.gov with the message **subscribe itl-bulletin**, and your name, e.g., John Doe. For instructions on using listproc, send a message to listproc@nist.gov with the message **HELP**. To have the bulletin sent to an e-mail address other than the From address, contact the ITL editor at 301-975-2832 or elizabeth.lennon@nist.gov.

The Computer Forensics Tool Testing Project

The sister project of the NSRL, the CFTT, is designed to provide a means for users to determine if a specific computer forensics tool meets their needs. NIST has taken on the task of defining a framework for testing computer forensics tools. The framework entails defining a means for

- classifying the functional characteristics and user requirements for different types of tools;
- specifying the details of these functional characteristics and requirements; and
- defining tests that can be used to determine if these tools meet the requirements.

Organizations can test the tools based on these requirements and the tests defined in the framework, and thereby have a measure of the capabilities provided by each tool.

As a side effect of this work, the computer forensics arena will also have a means for defining different sets of tool requirements in a standard framework. This framework will allow different subject area focus groups to work in specific domains without duplicating effort.

Initial work on the framework focused on the classification of tools by functional capabilities. Similar or related types of functionality were grouped to form broad classifications. The capabilities required in a classification were defined by focus groups of technical experts and practitioners that use specific types of tools.

There are two focus groups at present: disk imaging and write blocker. Other focus groups are planned in the future. The tasks of these focus groups are to produce a detailed specification of each functional classification and to review test assertions based on the specification. NIST then defines and prototypes test methods and test cases that can be used to exercise relevant functions of various tools purporting to provide the functionality required. At each step in the process, the set of requirements, test assertions, test cases, and procedures is distributed to subject area experts and posted to the CFTT website (http://www.cftt.nist.gov) for comment and peer review. Test results will be published by the National Institute of Justice.

Summary

The NSRL and CFTT are two projects based on collaboration among agencies to bring the effective and efficient use of information technology to the computer forensics arena. The NSRL is designed to save resources by automating tedious and labor-intensive components of computer forensics investigations. The CFTT is aimed at documenting tool capabilities in a rigorous manner. This should result in improving the quality of these tools. It will give the law enforcement community confidence in, and a measure of, assurance in the results of computer forensics tools used in investigations.

Disclaimer: Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.