2011-08-14

DEEP: Digital Evaluation and Exploitation Current Research

Department of Computer Science at the Naval Postgraduate School

Monterey, California: Naval Postgraduate School.

DEEP: Digital Evaluation and Exploitation
Department of Computer Science
Naval Postgraduate School
Monterey, CA

Current Research

From Domex

Jump to: navigation, search

Most of our current research is in the area of document and media exploitation (DOMEX) and computer forensics. Projects that we are currently working on in this area include:

Corpus Creation

We are creating multiple corpora for use in computer forensics research and education.

Test Data
: Test data is specifically constructed for the purpose of demonstrating a specific forensic issue or testing a specific feature in a tool. An example of this is our *nps-2009-ntfs1* test disk, which is a test image of an NTFS file system including unfragmented and highly fragmented files stored in raw, compressed, and encrypted directories (the decryption key is provided).
: Test data should be contain sufficient data to demonstrate or validate the specific item being tested, but, otherwise be simple and uncluttered with additional information. Test data can be freely distributed on the Internet without any controls.

Sampled data
: Sampled data is obtained by selecting a subset of a large data source, such as the Internet, using some kind of randomized process. The essential idea is to eliminate bias that may come from the use of a researcher's own data collection (*e.g.* documents or images from the researcher's personal hard drive). However, if true random sampling technique is employed, it becomes difficult to publish the set as it is impractical to ascertain that none of the files have any legal restrictions on redistribution.

Realistic data
: This data is similar to what a forensic investigator might encounter in an investigation, but the data set was in fact artificially constructed. Realistic data is typically created by performing clean installations of software on wiped machines. At this point the experimenter can run programs, perform basic operations, or even engage in sophisticated role play with other experimenters. Although there should be no privacy concerns when distributing realistic data, there may be copyright concerns.

Real and Restricted Data
: This data is created by actual human beings during activities that were not performed for the purpose of creating forensic test data. Access to this data should be controlled: it should not be placed on the Internet for anonymous download. Real data is typically subject to restrictions because of privacy or copyright concerns.

Real but Unrestricted
: These data sets can be (or have been) made available for unrestricted access. For example, the Enron Email Dataset is

a corpus of 619,446 real email messages from the 158 users inside the Enron Corporation. These email messages were entered as evidence in a court case by the US Government and, as a result, became publicly available without restriction. Another example of real but unrestricted data are photos that can be downloaded from the Flickr photo sharing website and user profiles on Facebook.

Our Test and Realistic datasets can be downloaded from:

- http://digitalcorpora.org

Hash Research

We have been engaged in work involving hash algorithms, including:

- The use of Bloom Filters for dramatically faster searching of hash databases, and for gauging the similarity between files.
- The use of sector hashes to identify residual data from files left on disks.
- Techniques to speed hashing using vector processors such as the IBM Cell Broadband Engine.

File and Fragment Identification for Rapid Drive Analysis

We have developed a fundamentally new approach for inventorying the content of a disk drive using statistical drive analysis. In order to be effective, this approach requires significant progress in solving the file fragment identification problem.

To date we have developed accurate file fragment identifiers for:

- JPEG
- MPEG
- Huffman compressed data
- HTML
- Encrypted data

Ascription

We are developing techniques for automatically determining the owner of indeterminate information found on a hard drive using other information on the hard drive as exemplars. We believe that we can apply this technique to:

- The Multi-User Carved Data Ascription Problem
- The task of determining network membership for the owners of portable storage devices.

Forensic Workflow

- Forensic Drive Analysis Process
- Large Scale Automation Analysis
- Retrieving SD Card Serial Number

[EDIT]
