High-Tech Holmes

By Jon Wright

Every crime has a crime scene that can be scoured for clues. But sometimes the evidence being analyzed isn't a bloodstain, a footprint, or a carpet fiber. It is the bits and bytes of data hidden inside a computer. In those cases, criminal investigators need to know how to coax secrets from the silicon chips. Much like their physical crime-scene counterparts, computer forensics investigators follow several basic steps. In the case of a crime that has already been committed, they must preserve evidence and then analyze what is collected. In the case of an investigation into an ongoing crime, they may be required to conduct surveillance of a suspect or a locale. And in either type of situation, they will ultimately have to prepare a detailed report of their findings.

Protecting Evidence

Digital data is easily destroyed. Therefore, when an investigator arrives at the lab with a computer in tow, the first priority is to preserve the integrity of evidence.

Even turning on a machine and booting up an operating system is enough to cause irreparable harm. For example, certain files have a date/time stamp that is updated when the system boots up. The date/time stamp that gets overwritten in the process would have shown when the user last turned on the machine--a potentially crucial piece of information. To avoid destroying evidence, forensic analysts will boot the computer up from a separate floppy disk or hard drive.

The next step, also designed to protect the integrity of the evidence, is to make a copy of what is on the computer or other storage media. All forensic analysis is conducted on these copies to ensure that the original data is not corrupted or altered in any way during the investigation.

To create the duplicates, forensic investigators use data-imaging tools that produce a bit-by-bit image of magnetic media, including hard drives, floppy disks, or high-density storage drives like Zip, and optical read-write media, such as CD-RWs. Though all forensic-imaging tools accomplish the same goal, they differ depending on the media and operating system they can image, how and where data is stored on the suspect computer, and the preferences and experience of the analyst.

Forensic data imaging tools make an exact copy of the entire contents of the medium, including hidden files and even deleted data. This type of copying is different from a normal backup of data, in which only the "logically accessible" files discussed below (such as Microsoft Word or Excel files) are copied.
In addition to imaging data, it may sometimes be necessary to recover and record data from damaged media, such as a CD-ROM that has been gouged. That's because suspects often try to destroy floppy disks by cutting them, smearing them with substances, dismantling the plastic casing, or twisting or crumpling the medium. In those cases, investigators can use custom software or hardware, available in some law enforcement laboratories and from some commercial sources.

Typically, forensic investigators copy data onto a medium that cannot be altered, such as a CD-ROM. However, it is not uncommon to encounter a 40-gigabyte hard disk, while a CD-ROM has a capacity of approximately 650 megabytes, so it is becoming increasingly difficult to copy data in this way. Some forensic facilities now use high-capacity RAID (Redundant Array of Independent Disks) drives for large-capacity storage. While precautions can be taken to prevent the alteration of data on a RAID, it is not a read-only medium like a CD-ROM.

Analyzing Evidence

Once investigators make an exact image of data, they can begin analysis. However, analysis is rarely easy or straightforward.

Among the billions of 1s and 0s stored on a hard drive or other media are operating systems, applications, files, utilities, e-mail, and more. Data may not be easily accessible, or it may contain booby traps or viruses. Even under the best of circumstances, the data that must be sifted through tends to be voluminous. Investigators tackle a haystack of data and work through it to find the one needle of evidence they need.

There are three broad and sometimes overlapping categories of computer data, each with its own particular challenges, that may need to be analyzed. Each type of data requires different tools for analysis.

First is logically accessible data (also called logical, or formatted, data), which can be especially challenging when encrypted, corrupted, or booby-trapped. If the data is in the form of images or sound or if there is an excessive amount of data to sort through, the difficulty of finding evidence increases. The second category of computer data is files that have been deleted, and the third is called unallocated space.

Logically accessible data.
Computers do not store files as one block of data. Rather, they scatter data across a magnetic medium, whether a hard or floppy disk, wherever there is space. Logically accessible data (the information stored properly on a magnetic medium, such as Word files, that can be accessed using normal techniques) has a table that lists the different areas on a magnetic medium where pieces of files reside. Thus a magnetic medium is like a book with an index but with scattered chapters and pages.

The index (called a File Allocation Table, or FAT, file) reassembles chapters and pages in the right order by using page number locators or "pointers." Logically accessible data is the easiest data to review, and the investigator will normally examine it first.

Analysts often speed the searching of logically accessible files by using software that allows some types of files, such as documents and graphic images, to be viewed without launching the specific applications used to generate and edit those files. Specialized string-search utilities are also useful. This software can search for particular keywords in the logical file structure, as well as in deleted files and file slack (more on this later).

Another timesaver is a utility called HashKeeper, which creates a numeric identifier called a "hash" from each file on a suspect computer. Each file identified by an investigator as not being evidence can be identified by its hash and excluded from future searches. Hashes also enable investigators to see identical files that have different names (for example, a suspect may try to conceal a graphics file called "photo.jpg" by renaming it "taxes.doc").

Encrypted data. Data encrypted by an individual can be the most challenging type of evidence to retrieve. Usually, investigators will not attempt to decrypt a file because it is an expensive and time-consuming process. Rather, they will look for evidence elsewhere in the computer that might reveal what is in the encrypted file. For example, there might be an unencrypted version of a file, or some traces of it might still remain in the system.

However, files that have been password-protected by applications such as Microsoft Word are sometimes easier to break. These applications often use a simpler type of encryption, and in many cases, the password itself can be broken using a password-cracking utility (whereas encryption software such as PGP can accept passphrases that are a sentence long and, thus, virtually impossible to crack).

Internet data. Data stored by a Web browser, which can be found in a favorites folder, a cookie, or the temporary history file, can indicate where a user has been recently. Investigators use specialized proprietary or widely available commercial tools to search browser caches for data that can reconstruct a surfer's history. In addition, various publicly available tools, such as WhoIs and Sam Spade, help the examiner track down the owner/operator of IP addresses.

Corrupted or booby-trapped data. It is now standard procedure for investigators to use virus scanners to search for malicious code embedded in evidence files before it can create havoc. Although most viruses reside in executable programs that will rarely be triggered during an investigation, newer virus types (specifically, macro viruses like Melissa and VBScript viruses like LoveLetter) exploit vulnerabilities present in some operating systems and applications and can be triggered simply by viewing the documents in which they reside. These viruses are potential threats to a computer forensics investigator. Few off-the-shelf tools can handle corrupted data. And traps can be extraordinarily dangerous. For example, a savvy criminal could rig a file that, when accessed incorrectly, could permanently erase data. Fortunately, they are rarely encountered. Nevertheless, experienced forensic analysts will work carefully to avoid any potential risks.

Image/sound data. Computers can contain thousands of images and sound files from the Internet and elsewhere, which pose further challenges to investigators, because these files can be masking information. For example, a process called steganography allows individuals to hide information by mixing it into the 1s and 0s that make up a picture. Government officials claimed recently that terrorist Osama bin Laden and his followers may be hiding plans and messages inside pornographic pictures on certain Web sites.

Forensic analysts can sometimes detect steganography if they are lucky enough to have access to the original file and compare it to the file believed to have been altered. More often the investigator will simply look for the presence of common steganography encoding tools such as S-Tools, Stash, JP Hide and Seek, and Scamdisk; if such tools were found on the suspect's computer, it is likely that they've been used.

Excessive amounts of data. Sometimes the issue isn't the type of data but the sheer volume of it. When data reaches into the billions of bytes, search utilities become unwieldy and searches grow beyond the capacity of a single computer. In one case, a forensics laboratory searched a year of a company's stored e-mail--as much as a trillion bytes (a terabyte) of data--for evidence of fraud. The only way to handle searching was through a customized solution. Investigators put together a "Beowulf cluster," a supercomputer built by interconnecting PCs using networking technology such as Ethernet and software programs that allow parallel processing. Custom-written software enabled them to filter through the e-mail and search for key terms.

Deleted files.
The second category of computer data is the deleted file. Deletion is a misnomer. When a file is deleted, its 1s and 0s are not erased. The computer simply removes from the index or FAT file the pointers to the file pieces scattered across the disk, thus making the space where the "deleted file" resides available for use by newer files. But until new 1s and 0s overwrite the 1s and 0s of the deleted file, it can be recovered. There are many commercially produced tools that can recover deleted files.

Unallocated space.
The third category of data is a catchall of items not fitting into the prior two. It includes files that cannot be recovered at all and unformatted data written to the media without a pointer, such as data held temporarily in holding bins called swap files or caches. (An operating system such as Microsoft Windows will temporarily dump data in a swap file while performing another activity and later retrieve it, allowing the system to manipulate files larger than its main memory.) Investigators look for data in unallocated space because it often contains evidence that can be found nowhere else.

Finding file slack and free space data usually requires its own type of software. One such program is called a carv utility, which locates files using information about the format in which the files are stored. These utilities come off-the-shelf, but many law enforcement and forensics firms use versions they have developed.

The carv utility takes advantage of the fact that most file formats (in particular graphic images) begin and end with specific bits of data that identify the start and end points for the file and its format. For example, all Web pages begin with <html> and end with </html>. The <html> at the beginning of the file can be considered the file header, and the </html> at the end, the file footer. A carv utility searches a drive's free space for known file headers, and it copies information associated with that header to a separate file. This process may locate individual files in free space even though there are no pointers to those files.

Carv utilities proved useful in a case involving a CEO whose former employer claimed that he had violated his noncompete agreement after going to work for a competitor. The ex-employer accused the CEO of trying to steal its clients, and it sought to prove that he intended to do so even before he resigned. The company believed that the CEO had received e-mails with attached spreadsheets from the competitor naming targeted clients. Those e-mails would help the company prove its case.

Forensic analysis of the CEO's computer found that, if the spreadsheets had ever been there, they had been overwritten and the pointers erased. However, because the spreadsheets had been sent attached to an e-mail message, they were in "encoded" format (e-mail attachments are specially encoded so they can be sent across the Internet). Those encoded files, written to a cache in the hard drive, were located by their headers and analyzed by forensic investigators. They showed that the CEO's new company had

targeted the clients that he already had a relationship with, indicating that he was being recruited to lure those clients away.

Surveillance

When a crime is suspected of being ongoing, the investigation takes on a different quality. It is no longer a question of a static crime scene with a set amount of evidence. Instead, it is a living scene with new evidence potentially being created daily. In such a case, forensic investigators may need to covertly use special monitoring tools to record the suspect's activities.

These tools, invisible to network users, keep activity logs to expose questionable usage patterns and security breaches. They can be triggered by any kind of user activity, application, or keystroke sequence (such as a visit to a certain type of Web site, the typing of certain words, or the use of applications such as Napster). An investigator can also program the software to monitor and record only unusual, unsolicited, or prohibited activities. These tools can help collect enough evidence to make a case. (Of course, laws regarding user privacy and other issues may pertain and must be observed.)

Findings

During a trial, investigators must present computer evidence in a logically compelling and persuasive manner that a jury will understand and an opposing counsel cannot rebut. This requires step-by-step reconstruction of investigators' actions with documented dates and times, and easily understandable charts and graphs that explain what was done and how. The result is testimony that explains simply and clearly what a suspect did or did not do.

When putting evidence into a useful format for presentation in a court of law, investigators must take the same care in presentation as they have in analysis. While forensic tool suites have formatting capabilities that make evidence building and reporting easier, many investigators find that piecing together the output manually using products like Microsoft Word, PowerPoint, and Excel results in a clearer and more comprehensive report.

Forensic challenges are wide-ranging, and the technologies to grapple with them are still developing. Because computers and networks change rapidly, forensic tools must do so too, and what is commonly used today might be retired tomorrow when a better tool appears. But the basics of investigations remain the most useful tool in helping the modern Holmes decipher the digital tracks of the modern Moriarty.

Jon Wright is the director of Computer Forensics and Incident Response for Veritect, Reston, Virginia.