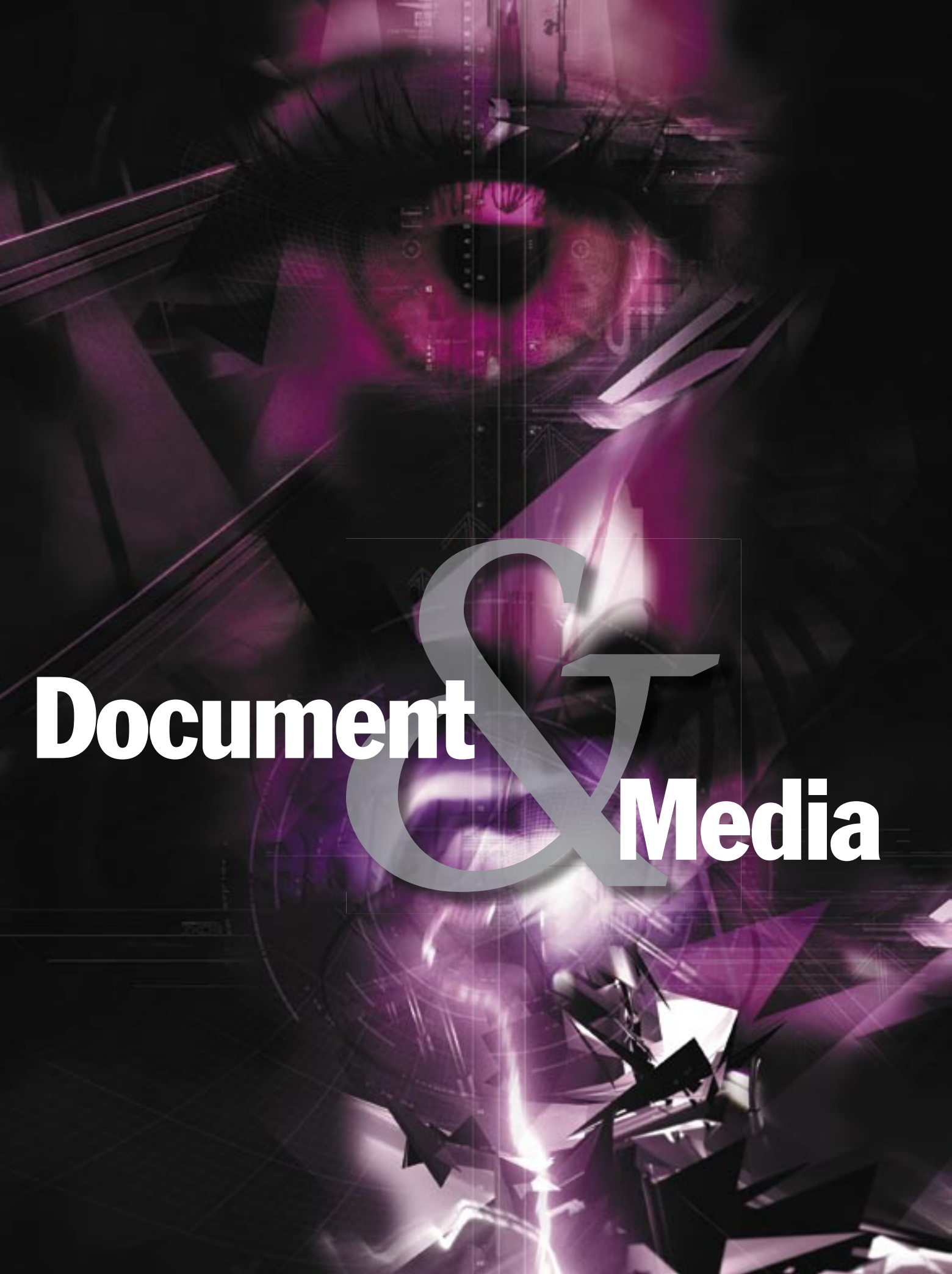
Document
&
Media

Exploitation

SIMSON L. GARFINKEL, PH.D.

The DOMEX challenge is to turn digital bits into actionable intelligence.

A computer used by Al Qaeda ends up in the hands of a *Wall Street Journal* reporter. A laptop from Iran is discovered that contains details of that country's nuclear weapons program. Photographs and videos are downloaded from terrorist Web sites.

As evidenced by these and countless other cases, digital documents and storage devices hold the key to many ongoing military and criminal investigations. The most straightforward approach to using these media and documents is to explore them with ordinary tools—open the word files with Microsoft Word, view the Web pages with Internet Explorer, and so on.

Although this straightforward approach is easy to understand, it can miss a lot. Deleted and invisible files can be made visible using basic forensic tools. Programs called *carvers* can locate information that isn't even a complete file and turn it into a form that can be readily processed. Detailed examination of e-mail headers and log files can reveal where a computer was used and other computers with which it came into contact. Linguistic tools can discover multiple documents that refer to the same individuals, even though names in the different documents have different spellings and are in different human languages. Data-mining techniques such as cross-drive analysis can reconstruct social networks—automatically determining, for example, if the computer's previous user was in contact with known terrorists. This sort of advanced analysis is the stuff of DOMEX, the little-known intelligence practice of document and media exploitation.

The U.S. intelligence community defines DOMEX as "the processing, translation, analysis, and dissemination

Document & Media Exploitation

of collected hard-copy documents and electronic media, which are under the U.S. government's physical control and are not publicly available."[1] That definition goes on to exclude "the handling of documents and media during the collection, initial review, and inventory process." DOMEX is not about being a digital librarian; it's about being a digital detective.

Although very little has been disclosed about the government's DOMEX activities, in recent years academic researchers—particularly those concerned with electronic privacy—have learned a great deal about the general process of electronic document and media exploitation. My interest in DOMEX started while studying data left on hard drives and memory sticks after files had been deleted or the media had been "formatted." I built a system to automatically copy the data off the hard drives, store it on a server, and search for confidential information. In the process I built a rudimentary DOMEX system. Other recent academic research in the fields of computer forensics, data recovery, machine translation, and data mining is also directly applicable to DOMEX.

This article introduces electronic document and media exploitation from that academic perspective. It presents a model for performing this kind of exploitation and discusses some of the relevant academic research. Properly done, DOMEX goes far beyond recovering documents from hard drives and storing them in searchable archives. Understanding this engineering problem gives insight that will be useful for designing any system that works with large amounts of unstructured, heterogeneous data.

WHY "EXPLOITATION?"

When researchers say that their work is centered on information or document "exploitation," eyebrows invariably raise. The word *exploitation* is provocative, attracting unwarranted attention to a process that could just as easily be classified as "computer forensics" or even "data recovery." But, in fact, the word is apropos.

The words *exploit* and *exploitation* imply using something in a manner that's "unfair or selfish."[2] And it's true. People who are in the business of document and media exploitation really do seek to make unfair use of computer documents and electronic storage devices. *Fair*, after all, means following the rules. The "rules" of a computer system are the APIs, the data-storage standards, the file permissions, and other interfaces that were intended to be used by the file's creator. When a file in the computer's electronic trash is deleted by "emptying the trash," the rules say that the file's contents should no longer be accessible. The "undelete" command that is part of every forensic toolkit takes advantage of the fact that computer systems generally do not overwrite the contents of deleted files. This is a common problem in computer systems, affecting not only deleted files in file systems but also deleted paragraphs in word processors and even unallocated pages in virtual memory systems.

Computer forensic practitioners working for police departments and litigation support firms also make their living by recovering intentionally deleted data, but even these processes follow rules—though those involved in exploitation might choose to ignore them. The goal of computer forensics is to assist in some kind of investigation, which usually begins because a crime was committed and, hopefully, ends with the perpetrator being convicted in a court of law. With conviction as a goal, forensic practitioners must be concerned with the evidentiary integrity and chain of custody—and they need to limit their search to information that is relevant to that investigation. In many cases the evidence will have been obtained under a search warrant or discovery procedure, the terms of which may limit the forensic examiner's actions or even which kinds of files may be examined. Evidence obtained by breaking the rules may even be suppressed.

For example, in the case of *U.S. v. Carey*, an investigator executing a warrant on narcotics discovered files with a JPG extension that contained child pornography. Carey was indicted and convicted for possession of child

pornography, but the appellate court reversed the ruling and remanded the case back to the trial court, arguing that "the seizure of evidence was beyond the scope of the warrant."[3] The evidence should have been suppressed.

Unlike the investigators in the Carey case, those engaged in document and media exploitation are not bound by any rules other than laws of physics and nature. The goal of information exploitation is to get and use the data—the ends justify the means. It's OK if these results aren't good enough for a conviction. Exploitation rarely seeks to prove or disprove the details of a case; instead, it seeks to make the fullest use of all the data that has been obtained. The standard of success is the usefulness of the result, not the reliability of the process.

If you find the preceding paragraph alarming, remember that DOMEX is about exploiting data, not people. "Exploitation" is precisely the attitude that you want when you take a crashed hard drive to a data-recovery firm. If you've just lost the only copy of a 400-page manuscript, it's probably OK with you if the firm is able to recover the first 200 pages of the September 20 version and the last 180 pages of the August 19 version. Although a good defense attorney might be able to suppress a document that was made by stitching together those two halves, you probably don't care about that if you are the author and the alternative is rewriting the 400 pages from memory. Likewise, if you are using some kind of desktop search system to index the files on your hard drive, you don't mind if the product makes a mistake or two and shows you files that you aren't "allowed" to see—just as long as you find what you're searching for.

THE TWO DOMEX PROBLEMS

Broadly speaking, DOMEX addresses two problems, which we will call "deep" and "broad."

The deep problem is the easier of the two to understand. Some kind of document or data-storage device—for example, a hard drive, DVD, or cellphone—becomes available for analysis. Because of the way that this object was obtained, we know that it is of interest. The goal is to find out everything possible about it.

A good example of the deep problem is the analysis of two hard drives stolen from Al Qaeda's central office in Kabul on November 12, 2001. The hard drives were in a laptop and desktop that Alan Cullison, a war correspondent working for the *Wall Street Journal*, purchased in Kabul.[4] Analysis revealed that the desktop had been used primarily by Ayman al-Zawahiri, one of Al Qaeda's top leaders. After Cullison verified that the computers were legitimate, he turned them over to U.S. intelligence officials. The analysts who were given those devices presumably wanted to know everything possible about them—not just the documents, but the application programs, the configuration settings, the other computers with which these machines had come into contact, and so on. Although few details of how these computers were analyzed have been made public, it would be logical to assume that every applicable forensic and document analytic tool in the U.S. arsenal was applied to the machines.

Another example of the deep problem is the analysis of a stolen Iranian laptop obtained by the U.S. government in July 2005. According to the *New York Times*, the laptop contained "more than a thousand pages of Iranian computer simulations and accounts of experiments" that "showed a long effort to design a nuclear warhead."[5] Once again, an analyst faced with examining this laptop would want to know everything about it that was technically and humanly possible.

The broad DOMEX problem flips things around. Instead of having unlimited resources to spend on a particular document, analysts are given a large number of digital objects and a limited amount of time to find something useful to an investigation. In recent years the amount of digital information seized during the course of law enforcement, intelligence, and even during civil litigation has exploded. "Ten years ago, a case would involve a few computer hard drives," e-discovery expert Jack Seward said in 2005. "Now a case is often hundreds of hard drives, numerous servers, and tape archives."[6] Indeed, a single case processed by the FBI's North Texas Regional Computer Forensics Laboratory in 2002 required more than 8.5 terabytes of storage and more than a month of computer work to process.[7]

This avalanche of digital media makes the broad problem quite compelling from both a national security and commercial perspective: a system that can reliably find the "good stuff" can save money, time, and perhaps even lives.

Document & Media Exploitation

Although these two problems may seem on the surface to be quite different, both require many of the same tools and technologies. Applying either approach to a hard drive requires software that can interpret disk structures for a wide range of operating systems and their different versions. The naïve way to do this is by mounting the disk partition read-only; a better approach is to use forensic file system recovery software such as The Sleuth Kit.[8] Such software knows how to decode on-disk file system structures, can recover deleted files, and is tolerant of data structures that might be missing or corrupt.

File recovery is just one of many required technical capabilities. Once files are recovered, software needs to extract salient "names and entities" such as human names, e-mail addresses, physical addresses, and so on. The software needs to be able to recognize variant spellings or codings for the same information. The system will probably need to build some kind of hypothesis about what kinds of processes inside the computer system created the stored data in the first place. Finally, the software must be able systematically to organize the information so that it can be automatically processed.

HUMAN-GENERATED CONTENT VS. TECHNICAL CONTENT

The intelligence community's emphasis on *translation*, *analysis*, and *dissemination* in its DOMEX definition is no accident. Much of the work on DOMEX grows out of previous work on DOCEX (document exploitation). Commercial DOCEX systems have been available to the U.S. government since the 1990s.[9]

Today there is still significant emphasis on documents, and on the information created by human beings. This is especially true when DOMEX information is presented in a criminal or civil trial. In a courtroom the prosecution can easily take a printout of an e-mail message or a digital photograph found on a hard drive and enter it into evidence. Certainly one reason that the Al Qaeda hard drive was valuable is that it contained correspondence with Osama bin Laden and other Al Qaeda leaders.

Technical content can be equally valuable. For example, an analyst might discover a link between two women because both women have photographs of the same man on their respective hard drives. Another way of discovering that link might be by determining that the two women both have digital photographs that came from the same digital camera (as identified by a serial number in an EXIF file) or because their copies of Windows XP were activated with the same stolen serial number. Information generated by a computer, such as a digital camera's serial number embedded in a JPEG EXIF record, might be critical in establishing a link between two individuals And unlike the analyst who recognizes the same man, the technical connection can be made automatically—even if the two hard drives are analyzed at two different locations—provided that there is a correlation step done at a central location.

Extracting technical information is complicated because many file formats are either proprietary or poorly documented. This kind of analysis is also rare among today's commercial forensic tools, which tend to focus on document recovery and presentation of data from a single drive. For example, a data-mining algorithm that discovers that an unprintable fragment of a PDF file has a common "ancestor" with another PDF file would probably not be useful in a court of law: explaining to a jury what such a match actually means would be difficult. Finding one of those PDF files on a captured laptop and another on a terrorist Web site, however, might be useful in helping an analyst understand how information flows through an organization.

The analysis of technical content is likely to grow more important in the coming years as the widespread availability of disk and file encryption makes human-generated content harder to access, just as the widespread use of encryption for communications increases the importance of traffic analysis for communications intelligence.[10]

AUTOMATED DOMEX

In the remainder of this article I present an architecture for performing automated document and media exploitation and show how the architecture can be applied to both the deep and broad problems of DOMEX. Although I use the example of a hard drive that arrives for exploitation, much of the discussion could apply equally well to a DVD or USB flash storage device.

STEP 1: IMAGING AND STORAGE

When a hard drive first becomes available for exploitation, its condition is generally unknown. The drive might be in perfect working order. On the other hand, the drive may have been damaged or about to fail and may have only a few minutes of operational life left. Therefore, when a drive arrives for exploitation, the drive's contents are typically copied to a high-capacity storage system such as a RAID or SAN (storage area network). This process is called imaging, and the tool to perform this task is called a disk imager.

A number of disk imagers have been developed for use by police departments and other computer forensic investigators. These programs make a sector-for-sector disk copy into one or more evidence files on the storage system.

Most forensic disk imagers will also calculate an MD5 or SHA-1 cryptographic hash of both the original disk and the image: by comparing the two the investigator can establish the faithfulness of the copy. In a criminal investigation this hash is recorded in a police or investigator's report; if the disk image is later provided to an expert working for the defense, that expert can verify that the disk image the defense team received is the same as the one acquired by the police.

A comprehensive list of disk imagers is available on the Forensics Wiki.[11]

The following additional features are desirable when a disk imager is used for DOMEX:

- The imager should be as automated as possible because of the potentially large number of disks that need to be processed.
- The imager should capture metadata about the hard drive such as its serial number, manufacturer, and firmware version, and handle bad sectors. (Some so-called bad sectors can nevertheless be read by turning off error correction; others can't. Some bad drives can even crash your host computer when you try to spin up the drive.)
- The imager should incorporate workflow automation features such as choosing a file name and storage location for the image file and detecting if the same drive has been inadvertently imaged before.
- In some applications, encrypting the image file with a public key may be desirable so the contents cannot be decrypted except in a secure facility.

Even though imaging is well-understood technology, many improvements are possible. Today's imagers need to be faster, more highly automated, and better able to handle disk errors. There is also a need for handheld imagers and covert imagers, as well as tools that can begin analysis before imaging is complete.

STEP 2: FILE SYSTEM ANALYSIS

A 60-gigabyte hard drive has 120 million 512-byte sectors, but thinking about the drive this way isn't terribly efficient or useful. Most hard drives have one or more partitions that may be one or more file systems. Each file system, in turn, has files that are *resident* and files that have been *deleted* but are nevertheless recoverable. This kind of extraction can be done with an open source tool such as The Sleuth Kit.

Once the file system metadata is extracted, it should be intelligently processed and stored in a common database. Documents can also be tokenized and indexed. Such a system makes it possible to rapidly search hundreds or thousands of disks by typing a single command.

After the disk's metadata has been extracted, a potentially large amount of data may nevertheless remain. This data comes from the sectors found between or at the end of partitions, sectors that were not ascribable to any file, and even bytes in the slack space at the ends of sectors and clusters. Forensic investigators who come up empty looking for incriminating information among the disk's files will typically use a carving tool such as Scalpel or Foremost to search through this additional space for digital images, word documents, and whatever other kinds of useful information they can find.

Although file-system analysis is a part of practically every civil and criminal forensic examination today, most of today's tools are designed for interactive analysis and do not work well in a batch environment. This is an area where research, engineering, and product development can have significant impact.

Another area where research is needed is in improving performance. Today's analysis tools, much like today's file systems, frequently rely on the head of the computer's hard drive for seeking information. Processing the contents of an entire hard drive (or hard-drive image) might involve a seek to every directory and then to every file—and that's before the carving starts. The problem here is that both disk capacities and data-transfer times are increasing much faster than the speed with which

hard drives can seek. As a result, a highly fragmented disk that can be imaged in an hour frequently might require 15 to 20 hours for the initial analysis—even if the image is stored on a high-performance SAN. A good research problem in the area of file-system analysis is the development of analysis software that operates in a streaming mode, reading the disk image from beginning to end and performing all necessary data analysis as the data flies by.

STEP 3: FILE ANALYSIS AND FEATURE EXTRACTION

Once the files are found, they need to be analyzed—automatically, if possible.

Today's computer forensic systems excel at document analysis, but only when used by a trained operator. Commercial "file filter" software is available that can understand, display on the screen, and extract the text from literally hundreds of different kinds of application data file formats. Once data is extracted, it can be processed with linguistic tools that can detect the language in which the document is written, translate the text into English (if necessary), or transliterate names and addresses into a standardized English spelling. The original language, the translations, and the transliterations can then all be stored in a full-text search engine, making it easy for a human analyst to rapidly search thousands of processed hard drives for a specific word or term.

Full-blown automated exploitation can go much further than simple indexing, of course. For example, hidden data from previous edit sessions is frequently left in Microsoft Word files; this data can also be automatically extracted and indexed.[12] Other information found in the metadata includes the time that edits took place and the registered name of the person performing the edits. JPEG image files record such details as the serial number of the camera that was used and the time of day; the JPEG format even has provisions for recording the GPS location of each photograph; logfiles found on virtually every hard drive can be used to build a network-centric map of the computer's electronic neighborhood. All of this information can be faked—but usually it isn't. Analysts can extract, archive, and exploit all this information.

For work that involved documenting privacy violations on discarded hard drives,[13] we wrote a program that could automatically find character sequences that had a high probability of being credit card numbers. Applying this program to a corpus of 150 hard drives, we could rapidly distinguish the few drives that had thousands of credit card numbers from the large number of drives that had hardly any. We then focused our investigation on these "hot drives." One of these drives had been used in an ATM before it was sold on eBay; another drive had been taken from a computer used for processing credit cards at a supermarket. Neither disk had been erased prior to being sold.

A surprising amount of both applied and basic research needs to be done in this area. Although some commercial and open source tools are available for data extraction, nearly all of them focus on extracting human-readable text rather than metadata that might be useful for secondary analysis. What's more, extraction software invariably lags behind the file formats used by commercial applications. For example, many open source programs can now process the OLE format used by Microsoft Word, Excel, and PowerPoint. Unfortunately, Microsoft is now moving to Office XML.

People engaged in criminal or terrorist activity may employ obsolete or obscure word processors, spreadsheets, and image file formats as alternatives to using encryption. This is because the presence of encrypted data may be a red flag, attracting the attention of an investigator. Data in oddball file formats, on the other hand, may simply be ignored by the average investigator unless there is reason to dig deeper. Thus, oddball file formats provide a kind of plausible deniability to those who are trying to hide the content of their communications.

Another research challenge is to develop automated software that can understand the data files on a hard

drive the first time they are encountered, without requiring someone to sit down and write a parser. Although this sounds like a fantasy, it really isn't. That's because the typical hard drive contains more than just data files—it also has the programs that process those data files. In theory, it should be possible to load those programs into a virtual machine, run them, and then have the programs read and process the document files. Many security researchers are now using this sort of approach for malware analysis. It should be usable for DOMEX as well.

STEP 4: ANOMALY DETECTION AND SOCIAL NETWORK ANALYSIS

At this point in the process, the data from the hard drive has been extracted, sliced, refactored, translated into multiple representations, and stored in multiple databases. Now the real work begins.

For the deep DOMEX problem, automated software should be able to perform an analysis that's at least as thorough as an analysis created by one or more humans. This is because the deep software can have access to a far greater store of forensic knowledge and techniques than even the most renowned investigator. Automated software, running with an appropriate database, can know practically every version of every program that has ever been sold commercially. It can create a detailed hypothesis of the ways that the suspect's hard drive must have been used, then look for additional evidence on the drive (or on the Internet) to support that hypothesis. Unlike an expensive forensic investigator, this automated software could be widely deployed within both the intelligence and law enforcement communities—assuming, of course, that someone would write it.

Automated software should also be able to excel at the broad DOMEX problem. A DOMEX facility that stores features from thousands of hard drives in a single database could perform large-scale correlations of features such as e-mail addresses or credit card numbers. This approach, called cross-drive analysis,[14] could determine if a particular hard drive was used by a person who has connections to a previously identified terrorist network. Alternatively, cross-drive analysis could be used to find a terrorist network in a sea of data from captured drives.

This database of the current information environment can also improve deep analysis. For example, finding scanned pages from an Al Qaeda training manual on a hard drive might be an important event—unless it's the manual that was discovered by the Manchester (England) Metropolitan Police and now resides on the U.S. Department of Justice Web site.[15] On the other hand, finding a file that matches the first 25 pages of the Department of Justice manual but then has divergent text might be exceedingly important.

STEP 5: REPORTING

Once the automated analysis is complete, the results need to be made available to others—investigators, analysts, or even the ultimate consumer of the intelligence product. Today these reports are created by human analysts who tailor the report for the needs and knowledge of the intended recipient. Not surprisingly, generating a report can be time consuming—sometimes more so than the actual analysis.

An automated DOMEX system could generate its own reports. These reports could be superior to current forensic reports, taking into account not just the subject material and the report's intended consumer, but also what information has already been reported to the consumer. That is, the DOMEX system could track each user's knowledge and fill in the gaps as necessary.

SEARCH AND RESEARCH

Each successive step in this hypothetical automated DOMEX system is further and further advanced from the current state of the art. Open source imaging, file extraction, and file-carving software are available from a variety of Web sites, but the reporting scenario described here is many years from being a usable technology.

Some civil libertarians have said they have reservations about the moral legitimacy of this work. Automated DOMEX systems, they fear, could easily become better surveillance tools for the masses. DOMEX software could be run covertly on desktop computers by large corporations, for example. Software that has the potential to be this invasive should not be developed, they argue.

Automated DOMEX software, however, actually has the power to improve privacy—not so much for the general public, but for people who are targets of investigation. Today there are far more disk drives to be analyzed than there are examiners to work with them. The result is delays that can both dangerously impede an investigation and damage the civil liberties of innocent suspects.

For example, in 2005 the United Kingdom passed legislation extending the time that terrorism suspects could be held without being charged from 14 days to 90 days, in part because the two weeks provided by the previous terrorism law did not provide sufficient time for the forensic analysis of a typical hard drive.[16] A high-confidence automated DOMEX system might give police the tools they need to clear a suspect in days, if not hours.

Document & Media Exploitation

CONCLUSION

As framed here, the DOMEX problem is very unstructured. You have a pile of data that intuition tells you is important. The challenge is to do something useful with it—ideally with as much automation as possible.

This kind of broad, unstructured problem makes scientists uncomfortable, since there is no hypothesis to test. It makes businesspeople uncomfortable because there is no obvious metric to measure success or failure. But this kind of unstructured problem dominates many of today's information-rich environments.

We have the data, but getting the data isn't the hard part—it's just the start. Q

REFERENCES

1. Intelligence Community Directive Number 302. 2007. Document and Media Exploitation (July 6).
2. *Oxford American Dictionaries*. 2005.
3. US v. Carey 98-3077, 172 f.3d 1268 (10th Cir. 1999).
4. Cullison, A. 2004. Inside Al Qaeda's hard drive. *The Atlantic Monthly* (September). http://www.theatlantic.com/doc/200409/cullison.
5. Broad, W. J., Sanger, D.E. 2005. Relying on computer, U.S. seeks to prove Iran's nuclear aims. *New York Times* (Nov. 13). http://www.nytimes.com/2005/11/13/international/middleeast/13nukes.html.
6. Kahan, S. 2005. Bring 'em back intact! *Accounting Today* (June 6-19). http://www.webcpa.com/article.cfm?articleid=13192.
7. Davis, M., Manes, G., Shenoi, S. 2005. A network-based architecture for storing digital evidence. In *Advances in Digital Forensics*, ed. M. Pollitt and S. Shenoi. IFIP International Conference on Digital Forensics, National Center for Forensic Sciences, Orlando, Florida (February 13-16).
8. http://www.sleuthkit.org/.
9. Document Exploitation (DOCEX) Transportable Support System (DTSS). 1997. *Commerce Business Daily* (July 31). Virginia Contracting Activity, MDA908-97-R-0040. http://www.globalsecurity.org/intell/systems/docex_dtss.htm.
10. Diffie, W., Landau, S. 1998. *Privacy on the Line: The Politics of Wiretapping and Encryption*. MIT Press.
11. Tools: Disk Imaging. http://www.forensicswiki.org/wiki/Tools:Disk_Imaging.
12. Byers, S. 2003. Scalable exploitation of, and responses to, information leakage through hidden data in published documents. AT&T Research. http://www.user-agent.org/word_docs.pdf.
13. Garfinkel, S., Shelat, A. 2003. Remembrance of data passed: a study of disk sanitization practices. *IEEE Security and Privacy* (January/February).
14. Garfinkel, S. 2006. Forensic feature extraction and cross-drive analysis. Digital Forensic Research Workshop, Lafayette, Indiana (August 14-16).
15. http://www.usdoj.gov/ag/manualpart1_1.pdf .
16. House of Lords and House of Commons Joint Committee on Human Rights. 2005. Counter-terrorism policy and human rights: terrorism bill and related matters. Third Report of Session 2005-06, HL Paper 75-I, HC 561-i. http://www.publications.parliament.uk/pa/jt200506/jtselect/jtrights/278/27802.htm.

LOVE IT, HATE IT? LET US KNOW

feedback@acmqueue.com or www.acmqueue.com/forums

SIMSON L. GARFINKEL is an associate professor at the Naval Postgraduate School in Monterey, California, and a fellow at the Center for Research on Computation and Society at Harvard University. His research interests include computer forensics, the emerging field of usability and security, and personal information management.

The views expressed in this article are solely those of the author and do not necessarily reflect the positions or policies of the Naval Postgraduate School or the U.S. Government. This article describes the author's research and is issued to further discussion.