

TECHNOLOGY A Special Report

Copy, Paste, and Reveal

Electronic redaction doesn't always hide what it's supposed to hide.



By **DANA J. LESEMANN**

With the issue of intentional government leaks of classified information frequently in the news, the problem of unintentional leaks of classified and sensitive information is frequently overlooked. The examples are numerous and startling.

Last April, U.S. military commanders in Iraq released a long-awaited report of the American investigation into the fatal shooting of an Italian agent escorting a freed hostage through a security checkpoint. In order to give the classified report the widest possible distribution, officials posted the document on the military's "Multinational Force-Iraq" Web site in Adobe's portable document format, or PDF. The report was heavily redacted, with sections obscured by black boxes.

Within hours, however, readers in the blogosphere had discovered that the classified information would appear if the text was copied and pasted into Microsoft Word or any other word-processing program. *Stars and Stripes*, the Department of Defense newspaper, noted that the classified sections of the report covered "the securing of checkpoints, as well as specifics concerning how soldiers manned the checkpoint where the Italian intelligence officer was killed. In the past, Pentagon officials have repeatedly refused to discuss such details, citing security concerns." Soon after, the report was removed from the Web site.

Copies of the improperly redacted report, however, live on. We at the consulting firm of Stroz Friedberg, too, were able to remove the redaction and save the clear text in a Word document. Forensic examiners in our office found that the document had been produced directly from Microsoft Word using Adobe Acrobat 6.0's PDFMaker. The redacted text simply had been highlighted in black. As a result, to reveal the classified information, the steps are simple: Highlight the text with the "select text" button on the PDF toolbar, copy the text by typing "control C," open a new document in a word-processing program, and paste the text into the new document.

This is not an isolated incident. During the last presidential campaign the Department of Defense posted on a DOD Web site 22 documents and one photograph concerning President George W. Bush's service in the Texas Air National Guard. The photograph and one of the documents are in jpg format; the rest are in PDF format.

Redactions of personal identifying information have been made to almost all of the documents. In some cases it is clear to the naked eye that the redaction has been made by hand, with the traditional black marker on paper. In other cases, however, the neat, clear lines of extraction indicate that a computer program has been used to black out the text.

Using free and publicly available software, Stroz Friedberg's forensic examiners converted two of the electronically redacted PDF documents on the Web site to Word documents: a memorandum dated July 22, 1971, listing the members of 1st Lt. George W. Bush's Fighter Interceptor Squadron, and a form titled "Request and Authorization for Active Duty/Active Duty for Train[sic] Full Time Training (ANG)," dated June 25, 1973. When the conversion was completed, the redaction disappeared and the addresses and Social Security numbers of Bush and 10 of his colleagues appeared. According to the metadata in these documents, the PDFs of these two documents were created on Oct. 6, 2004, six months after the Web site was created.

After we notified the Department of Defense of this problem, the department apparently printed out the redacted documents from the Web site, then re-scanned them into the Web site as an image and not as text, which ensured that the redacted information could not be restored.

This is not a new problem. *The New York Times* ran into a similar problem in redacting data in June 2000, when it published PDF files of a classified CIA report titled "Overthrow of Premier Mossadeq of Iran" on its Web site. The *Times* noted that some names and identifying descriptions had been removed from the report because "there might be some serious risk that

the families of some of those named as foreign agents would face retribution in Iran.”

But John Young, the administrator of the Web site www.Cryptome.org, which specializes in publishing national security documents, was using a slow computer to view the *Times*' Web site. He noticed that the full text of the document appeared briefly on the page before the redactions appeared. The paper had not removed the names and descriptions but had only obscured them with black lines and boxes. By interrupting the page from loading, he could transcribe a portion of the hidden names, which he e-mailed to the *Times*. The following day the *Times* removed the report from its Web site.

Uncovering redacted information from the 2002 indictment of Gary McKinnon was even simpler. McKinnon, a British national, was indicted in the Eastern District of Virginia and the District of New Jersey for allegedly breaking into 97 computers belonging to the U.S. Army, Navy, Air Force, Department of Defense, and NASA, as well as six computers belonging to three universities, two public libraries, and a private business. The indictment described the computers McKinnon had allegedly hacked, including the actual IP addresses electronically redacted. The IP numbers may appear, though, if users highlight the redacted text and copy it into a new document. The redacted addresses will appear in the new document.

An IP address is a unique number, like a telephone number, used by computers on a network to identify route information and communicate across the Internet. IP addresses are assigned to all computers that connect to the Internet using the TCP/IP protocol. These include the basic home computer, Web servers, and firewalls protecting users of an organization or company. When surfing to a Web site on the Internet, a user types the name of the site, but the computer identifies that site by its unique IP address.

The IP addresses listed in the McKinnon indictment were all from a block of IP addresses assigned to the U.S. military. The block assigned to the military was not secret or unknown, but what McKinnon did was scan every number within the block, identify the computers that were vulnerable, and publish that list to the Internet. This is like a burglar trying every phone number in the 202 area code and then publishing a list of those individuals who answered and said they leave their front door unlocked.

By failing to properly redact these IP numbers in the indictment, the Department of Justice essentially republished the addresses of the vulnerable computers. When some of us at Stroz Friedberg alerted the DOJ to the problem, personnel there said that the publication of the IP addresses was of no concern because the information had been made public in another document.

WHY DOES THIS HAPPEN?

Why do these leaks keep recurring? Why do apparently well-meaning individuals, who are trying to distribute information widely and believe they are properly redacting classified information, repeatedly compromise national security when they publish information electronically in PDF? The answer is that most people simply do not understand how PDF files work.

Before the information age, document distribution was accomplished by hand. Duplicating the document via photocopying and physically removing sensitive sections was tedious but generally secure.

With the advent of the computer and electronic technology, image scanning became possible. Digitized documents meant infinite distribution and accessibility. The tagged image file format (TIFF) soon became the format of choice for scanning and distributing documents for three reasons. First, TIFFs have the capacity to save multiple distinct pages in a single file. Second, documents can be stored at different levels of compression and image quality, depending on file size requirements. This quality is commonly referred to as “scalability.” Third, TIFFs do not have associated metadata such as the author of a document or dates the document was created and last modified, which could inadvertently reveal sensitive information.

But TIFFs are scanned images and thus have one inherent flaw: They have no text editing, searching, or copying functionality. The next large step in the ability to handle documents came from digitizing documents. Optical character recognition (OCR) extracted text from images in a digital format. Scanning documents into a digital format preserved their layout and allowed for text searching. Until Adobe created the PDF, however, there was no format that combined text and images and allowed users to search both.

The creation of PDF files, therefore, was a giant leap forward. PDF files offer simple and critical features for working with documents while leaving complicated, size-increasing, unnecessary features for others to provide. PDF files can be made from scanned images, digital photos, other PDF files, text documents, or any printable file on a computer. Essentially any file that has visible content can be converted into a PDF file. PDF files allow restrictions on document viewing, printing, editing, content copying or extraction, adding comments, changing the field fills or signatures, content access, and document assembly. They also allow customizable password protection and encryption.

In addition, people appreciate the simplicity of the free and accessible programs for viewing PDF files. The availability of free PDF viewers means that large institutions, including corporations, universities, and government agencies, can produce documents and make them accessible to anyone with a computer while limiting the printing and distribution costs associated with hard copies.

The key difference between a PDF and an image is that PDF files can contain layers, including form fields, such as fill-in blanks; individual objects, like pictures; and, most important, fully functional text that can be searched, highlighted, copied, and pasted. These layers, which provide PDF files with their functionality, also hold the key to understanding the errors in redacting PDF files.

When an editor redacts text with a black marker, the text is not gone but simply hidden under a layer of ink. Similarly, when an editor redacts text in a PDF file, the editor is simply placing one PDF layer over another: A black box is placed over the text to be redacted. If nothing else is done, a reader can either move the black box or move the redacted text out from under the redacting object.

For its part, Adobe does not claim that the tools that come standard with Acrobat can be used to redact classified material. In fact, Adobe's marketing materials suggest using vendors to "permanently hide confidential information located in a document." Stroz Friedberg's forensic examiners tested one vendor and found that, in fact, it did permanently redact information and maintained the PDF file's functionality.

So when we called Adobe's customer service for information about redacting information from PDF files, we expected to be referred to an outside vendor. Instead the customer service representative suggested we do exactly the same thing that had created the problems with President Bush's records: Draw a black box over the text to be redacted, then fill in the box so the text cannot be read. The representative was apparently learning as he went along. As he performed the steps he realized that the method he prescribed would not work: The text box could easily be removed. He then suggested drawing a black rectangle over the text, saving the document as a TIFF file, and creating a PDF from the TIFF.

The customer service representative is correct: This method is absolutely secure—once the document is a TIFF, the redaction cannot be removed. But the document also no longer has any of the properties that make a PDF so attractive; as a TIFF, it no longer has text functionality without using OCR software.

The simplest way to ensure that information is secure is never to distribute it to anyone. But, of course, that is no solution.

THE SOLUTION

Individuals at all levels of an organization must understand the technology being used before it is deployed, especially when dealing with classified and sensitive information.

The increasing use of Adobe PDFs across the public and private sectors is a step forward because it leads to a wider dissemination of information. This growth also represents new risks, however.

Lawyers, and anyone else who works with privileged and classified information, must understand the technology that they are using before it is deployed. Thus, law firms, corporations, government agencies, and the military all must implement best practices for redacting sensitive data from electronic documents. This is not a matter of simply purchasing software designed for redaction, installing it on some machines, and training a few people. A law firm's managers must understand what trade-offs they are making in terms of the wider dissemination of information at the cost of security. The firm's IT group must ensure that the proper checks are in place and that they work to protect sensitive information. Finally, paralegals and other professionals who review and redact information must understand the steps they must take to protect the information in the appropriate forms.

Until then, law firms and other organizations that deal with the redaction of sensitive information should strongly consider returning to the days of the black marker and the TIFF. The TIFF is certainly not as consumer-friendly as Adobe's format, and it is not the latest technology, but at least it is secure.

Dana J. Lesemann is vice president and deputy general counsel of Stroz Friedberg, a consulting firm specializing in computer forensics, with offices in New York, Washington, D.C., and Minneapolis. Don Allison and David Morgenstein of Stroz Friedberg contributed to this article.