Computer Forensics Technologies for Personally Identifiable Information Detection and Audits

Yin Pan, Ph.D., is an associate professor in the department of networking, security and systems administration at the Rochester Institute of Technology (RIT) (New York, USA). Pan is actively involved in teaching and research in the security area, especially in IT security audits and computer forensics. She has published many papers in these fields.

Bill Stackpole is an assistant professor in the department of networking, security and systems administration at RIT. His teaching has focused in the areas of system administration, system security and computer system forensics. He has additional research interests in the areas of authentication and virtualization.

Luther Troell, Ph.D., is a professor in the department of networking, security and systems administration at RIT. Troell currently serves as the director of the school of informatics and teaches networking and information assurance classes. He has presented papers on curriculum development in information assurance as well as the development of an undergraduate program, Information Security and Forensics, and a graduate program, Computer Security and Information Assurance.

Identity theft has become more prevalent in recent years; about 10 million incidents occur each year.[1] IT professionals must understand the need for personally identifiable information (PII) discovery to protect themselves and their company from the civil, legal and financial liabilities caused by data loss. As documents migrate to digital form from hard copy, sensitive personal information gets stored in a variety of places digitally. National and international laws are in place requiring companies to search for confidential data to ensure compliance. Some US examples include the Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act (HIPAA). At the state level in the US, New York State's Disposal of Personal Records Law (2006) requires businesses to "properly dispose of records containing personal information," implying that this information must be unreadable and unrecoverable. International privacy laws, many of which are more stringent than those in the US, require similar activity.[2]

To comply with these laws, security professionals use a variety of sensitive information discovery tools to find and remove readily available information stored on end-point devices. While current PII discovery tools can find information that is readily available, they are not capable of discovering information that has been encrypted, obfuscated, hidden, deleted or is otherwise unrecoverable. It is critical to note that the content and metadata of deleted files can be easily recovered using standard forensics tools.

This paper will introduce computer forensics techniques to reveal sensitive data that are likely to be missed by PII tools, including data in RAM memory, graphics files, registry information or files marked as deleted.

PII BACKGROUND

Where Do Existing PII Tools Look for Sensitive Data?

PII tools typically search through the directory tree, looking at the content of allocated files, e-mail and the like for keywords or strings that indicate the file needs further investigation. For example, the tools might search for files that contain strings of 16 digits that could represent a credit card number. The strength of current PII tools is to quickly find possible PII in locations in which data are visible to the operating system. However, they do not provide support for searching data outside of these traditional areas. For example, if a file containing the same string of 16 digits had been deleted, the current PII tools would not find the file even though it may be recoverable. Deleted files are considered unallocated and are not available to view by PII tools.

Where Else Can Information Reside and Hide?

Data can reside in places that the operating system can see as well as places that are invisible to it. Most of these nontraditional locations are completely overlooked by PII tools but can still harbor sensitive information. Some examples of these areas include:

- **Metadata**—Metadata contains file information about the actual data in the file. Metadata can include who created the information, what the information is about, when the file was created/modified/accessed, and other information that can describe particulars about the data file. Metadata can provide a place for sensitive data to reside, but it is not particularly easy to remove.
- **Recycle bin and unallocated space**—When a file is deleted in a Windows machine, for example, the file appears in the recycle bin. The complete path and file name are stored in a hidden file called INFO2 in the recycled or recycler folder.

After emptying the recycle bin, the corresponding entries in INFO2 are also deleted. However, when the file is removed from the recycle bin, the system simply marks the space previously occupied by this file to be "available for use." Metadata information as well as the original contents of the file remain on the hard drive, completely intact until the space is allocated to new files and the original information is overwritten with new data. As storage becomes less expensive and hard drives increase in size, deleted files are likely to remain in unallocated space longer.

- **Alternate data streams**[3]—Microsoft's New Technology File System (NTFS) uses "streams" to store file content. A stream is a link to a data storage location. NTFS allows multiple streams of data to be associated with a file but only the default (main) stream is normally displayed to the user. One could hide sensitive data in any number of alternate streams associated with a host file with no apparent changes to the host file.
- **Files with modified extensions**—File extensions are the characters following the dot in a filename, e.g., text.txt. They indicate a file's data type but can easily be changed by the user. Each file also has a file header that uniquely identifies the file type. Adversaries often attempt to disguise the true nature of a file by changing the file's extension. Such changes can make it difficult for PII detection tools to find sensitive information in a file.
- **Graphics and images**—Pictures, images and graphics can contain information that may be considered private or sensitive. For example, a screen capture of a spreadsheet page with account or other character-based information will not be read as a character-based file.
- **Print spool files**—Printing involves a spooling process. For a print job, the file's content is written to a spool file (.SPL) and a separate graphics file (.EMF) for each page. These Print Spool files are saved to disk until the print is done and then deleted.
- **Link or shortcut files**—Link files are shortcuts pointing to the actual files that allow users to launch programs, open files and folders, or connect to a URL. They contain path information to allow the operating system to navigate to the location of the actual data file.
- **RAM and page files**—Some sensitive information may be found only from the computer's RAM and in its page files. At other times, such data will be written to disk if a machine is suspended or hibernated.
- **E-mail**—E-mail, instant messages and other communication traffic can contain sensitive information.
- **Registry hive**—The Windows registry contains settings specific to the hardware, applications, services, security and users on the system. Much potentially valuable information is stored there. Such information includes user and group identities and passwords and Internet history, including cookies and records of queries such as searches and lists of recently accessed files.

PII TOOLS AND FINDINGS

Existing PII tools

Both open source and commercial products are available to assist in detecting sensitive data on digital media. Some examples include:

- Find_SSN
- Sensitive Number Finder (SENF)
- Spider
- Identity Finder

Find_SSN is a simple-to-use Windows application. Developed by Virginia Tech, Find_SSN searches for files that may contain data matching the known patterns of Social Security numbers. Find_SSN does not scan PDF and Microsoft Outlook PST files, and files larger than 100 megabytes. SENF is a tool written by the information security group at the University of Texas at Austin. It has similar search capabilities as Find_SSN. Cornell University's Spider is an open source, more sophisticated scanning tool that can scan archives, documents and spreadsheets after selecting a target directory. However, Spider does not scan PDF and Microsoft Outlook PST files. Identity Finder is a commercial search tool capable of locating and identifying personal information (such as Social Security, credit card and bank account numbers as well as passwords or dates of birth) in files, e-mail, databases, registry entries and web browser caches. However, even with Identity Finder, there are still many areas not in its search path.

Testing the PII tools

To test the limits of the existing PII tools, a variety of files were created including .pdf, .xls, .doc, .txt and archive files on a USB drive. Two of these files were deleted to test whether the tools search for deleted files in unallocated space. The file text.txt was also renamed to text.jpg to test whether the tools are capable of detecting sensitive data from a file with

its extension modified. A file was also created that did not contain Social Security numbers in its main content, but had these data hidden in an alternate data stream. Note that none of the deleted files had been overwritten.

In addition, the tools were run against a live Windows machine so they could interrogate files that could not be stored on the USB drive, such as deleted Microsoft Outlook (*.pst) files, the Windows registry, alternate data streams, files in the recycle bin, print spool files and RAM data.

To interrogate files from the USB test drive, each tool was directed to run against the drive image and the results were recorded. For other files not stored on the drive, the tool was directed to the location on the hard drive where those files would normally reside. All the test files are listed in **figure 1**.

Results From Running These Tools

The Find_SSN, SENF and Spider tools discovered only a limited subset of the conditions presented. Identity Finder, on the other hand, improved on the capability to identify PII when compared to the other PII tools tested. Identity Finder not only detected all of the conditions found with the other tools, but also discovered PII in renamed files, PDFs, registry entries and .pst files. However, all of these tools still had limited capability to detect sensitive information stored in metadata, alternate data streams, graphics files, printer spools, RAM and page files. Additionally, none of the tools could identify PII that exists in unallocated space (if the recycle bin has been emptied, for example) even though the PII remains intact on the disk. All of these areas represent potential data leakage areas for the current PII tools.

A summary of the results is shown in **figure 1**.

Limitations of PII Tools

The current PII identification tools search for sensitive data—either through user-specified directories or starting at the root directory. They are not designed to find information that has been obfuscated, encrypted, deleted or otherwise hidden. **Figure 1** shows that PII identification tools miss finding information in areas including unallocated space (i.e., deleted files that can be recovered), e-mail and deleted e-mail, files open and locked or auto-saved by other processes, system files (executables, DLLs, hibernation files), the Windows registry, application-specific database files, memory and page files, metadata, graphics files (screenshots, thumbnails, RPM, TIFF, JPG, EMF, etc.), intentionally hidden PII in digital data (digital steganography, for example), as well as links or indirect references to PII. These limitations may lead to inaccurate results with false positives and false negatives.

FORENSICS TOOLS AND FINDINGS

As mentioned previously, the current existing PII identification tools search only within the file system and are not capable of detecting information in deleted files, or from any location not normally accessible to a nonprivileged user. Modern forensic tools can help to bridge this gap. Forensic tools are designed to recover and analyze data that has been intentionally or unintentionally deleted or otherwise hidden. Originally, this class of tools was associated with activities in the law enforcement sector and was used exclusively to discover legal evidence. Recently, they have been seen widely used in business and corporate environments.

Forensic tools are capable of bypassing the limits imposed by the operating system and can find file content that has been deleted (i.e., no longer available to the operating system) or has been stored in a place not typically accessible to a user, such as RAM or the registry. They can display files stored in a variety of formats, allowing a knowledgeable user to find information that would otherwise appear to be inaccessible.

While forensic tools' strength is their capability to search nontraditional areas in which data can be hidden, their weakness is that they can be time-consuming as they search so much more of the system than PII tools.

Existing Forensics Tools

While many forensics tools may be capable of detecting PII, two commercial forensics tools, Forensic Toolkit (FTK)[4] and EnCase,[5] were studied.[6]

Both EnCase and FTK run on Windows and provide sophisticated digital evidence analysis functions such as recovering deleted files, keyword and regular-expression searching, registry viewing, e-mail and memory analysis, and much more. The popular, versatile and free forensics tool the Sleuth Kit (TSK) with an advanced interface, PTK,[7] can also perform most of the same tasks. These tools can be used on offline as well as running systems.

Finding PII Using Forensics Tools

The forensic tools were applied to interrogate the same dataset tested by the PII tools. The results were then incorporated into **figure 1**.

Figure 1—Test Files and the Results

File Name	Description of the file	FindSSN	Spider	SENF	Identity Finder	FTK	EnCase
My Recent Documents	A folder containing a link file that links to text.txt	No	No	No	No	FOUND by following link	FOUND by following link
SSN test file.pdf	A PDF file containing Social Security numbers	No	No	No	FOUND	Viewable *	Viewable*
Text.jpg	A text.txt file, renamed to a JPEG file	No	No	No	FOUND	FOUND	FOUND
Earnings.xlsx	Excel spreadsheet	FOUND	FOUND	FOUND	FOUND	FOUND	FOUND
PII detection test.ppt	PowerPoint slides	FOUND	No	No	FOUND	FOUND	FOUND
SSN test file.docx	Word document with Social Security numbers in content	FOUND	FOUND	FOUND	FOUND	FOUND	FOUND
SSN test file-deleted.docx	Word document with Social Security numbers in summary (metadata)	No	No	No	No	FOUND	FOUND
	It was printed to generate a print spool file (*.emf)	No	No	No	FOUND	Viewable *	Viewable*
	Deleted Social Security numbers test file	No	No	No	No	FOUND	FOUND
ScreenShotWithSSN.png	A screen shot containing Social Security numbers	No	No	No	No	Viewable *	Viewable*
textFileCyptoClass.rtf	RTF	FOUND	FOUND	FOUND	FOUND	FOUND	FOUND
Text.txt	Text file	FOUND	FOUND	FOUND	FOUND	FOUND	FOUND
Text-deleted.txt	Deleted text file (after recycle bin emptied)	No	No	No	No	FOUND	FOUND
PII test.zip	Zip file containing text.txt and Social Security number test file.docx	FOUND	FOUND	No	FOUND	FOUND	FOUND
pst file	Outlook file with e-mail not deleted	No	No	No	FOUND (limited support)	FOUND	FOUND
deleted pst file	Outlook file with e-mail deleted	No	No	No	No	FOUND	FOUND
File with SSN in alternate stream	Word document with Social Security numbers in alternate data stream	No	No	No	No	FOUND	FOUND
File in recycle bin	File deleted but recycle bin not emptied	FOUND	FOUND	FOUND	FOUND	FOUND	FOUND
RAM and page files	Contents of memory with Social Security numbers in memory	No	No	No	Unknown	FOUND	FOUND
Windows registry	PII written to Windows registry	No	No	No	FOUND	FOUND	FOUND

* While the tools are not capable of searching these files directly, they allow display of the enclosed image using gallery view.

Both Encase and FTK offer live and static data search functions that support pattern matching using regular expressions. For example, a search might be conducted to identify files containing Social Security numbers using the pattern: <\d\d\d[\-]?\d\d[\-]?\d\d\d\d\> (where \d represents a "digit"). Using this search, the forensic tools discovered Social Security numbers from all deleted files, the signature mismatch file, the ppt file, RAM and page file, e-mail including deleted e-mail, link files, and registry hives. Data that had been hidden in the alternate data stream were revealed as well as the PII stored in the metadata of the file. These are not trivial differences. A file containing PII in its metadata would not be discovered by any of the previously discussed PII tools, since they focus only on file content. Using a process called signature analysis, the PII hidden in a file with its file extension changed from .txt to .jpg is also easily discovered by the forensics tools. The signature analysis process identifies and corrects the changed extension by comparing the file extension with header information stored in a file.

Even though the powerful live search reveals many results that were missed by the PII identification tools, the sensitive information hidden in graphics files, a print spool file and a PDF file were not detected by live search. The authors also utilized other technologies (built into the forensic tools) to uncover the rest of the PII information:

- **Graphics files**—To identify sensitive information hidden in graphics files, the authors used the Gallery/Graphics View feature. By clicking the Graphics/Gallery View tab, all graphic formats including EMF, BMP, TIFF, JPEG, PNG and GIF are displayed. The operator can then view the files to make a determination about the sensitivity of the image content.

  It may be possible to convert image files to text using optical character recognition (OCR) techniques. Regular-expression searches can then be employed to search for strings in the resulting files. This capability is not currently a feature of for-free or for-pay forensic tools, but such capabilities could improve PII scanning exercises as well as the forensics tools.
- **Print spool files**—Forensics tools support data carving that allows an investigator to search for EMF, GIF, JPEG, PDF, HTML and MS Office document files embedded in other files or unallocated space. This feature can recover deleted files embedded in unallocated space as well as the EMF files embedded in print spool files. Even though the EMF files are deleted after the print is done, forensics tools are capable of recovering deleted EMF files from unallocated space. Once the EMF file content has been recovered, the information is then viewable using the forensics Graphics View feature described previously.
- **PDF files**—Regular-expression live search does not support searching for PDF files. PDF files must be converted to text using OCR techniques for the live search. For this research, the authors viewed all PDF files via EnCase/FTK's Transcript/Native View tab. This is not a viable solution if there are many PDF files to be searched. Deleted PDF files can be recovered using the data carving feature.

In summary, using the forensic tools, all of the traditional locations for PII, as well as many areas in which PII could be intentionally or inadvertently hidden, can be searched. Forensics tools can add significantly to the results provided by PII tools.

CONCLUSIONS

A fundamental strength of forensics tools when used for detecting sensitive information is that they can recover deleted files and embedded images, list all alternate streams for each file, and perform signature analysis before conducting a search for sensitive information. A search using forensics tools not only checks the file content, but also examines file slack, metadata, links and other content. As a result, forensics tools are capable of discovering more information than the PII detection tools that were evaluated. They provide an additional source of information by which a PII search can become more effective and uncover more potential sources of PII.

However, there is no single tool capable of effectively finding all PII. Each of the tools has its strengths and weaknesses. This article was intended to address data that are likely to be missed by currently available PII tools. Forensics tools can be a powerful addition to the auditor's toolbox, providing additional capabilities to reveal sensitive data. If an auditor uses only PII tools, it is possible that sensitive recoverable information will be overlooked.

Organizations concerned about unauthorized dissemination of sensitive information should be aware of the limitations of their current PII tools and use this knowledge to make decisions about potential data leakage or compromise. Otherwise, it might be interpreted as a failure to adhere to privacy laws or regulations that could subject the organizations to legal liability or negative publicity.

ACKNOWLEDGEMENT

The authors wish to acknowledge the reviewers for their invaluable comments and suggestions that have improved the overall quality of the paper.

ENDNOTES

[1] SpendonLife.com, "2009 Identity Theft Statistics," *www.spendonlife.com/guide/2009-identity-theft-statistics*

[2] Information Shield, "International Privacy Laws," *www.informationshield.com/intprivacylaws.html*

[3] Parker, Don; "Windows NTFS Alternate Data Streams," *Security Focus*, 16 February 2005, *www.securityfocus.com/infocus/1822*

[4] Forensic Toolkit (FTK), *www.accessdata.com*

[5] EnCase Forensic, *www.guidancesoftware.com*

[6] The authors do not endorse these products in any way.

[7] PTK and Sleuth Kit, *http://ptk.dflabs.com*