Signature analysis and Computer Forensics

Michael Yip
School of Computer Science
University of Birmingham
Birmingham, B15 2TT, U.K.

26th December, 2008

Abstract: Computer Forensics is a process of using scientific knowledge to collect, analyze and present digital evidence to court or tribunals. Since files are the standard persistent form of data on computers, the collection, analysis and presentation of computer files as digital evidence is of utmost essential in Computer Forensics. However, data can be hidden behind files and can be enough to trick the naked eye. Therefore, a more comprehensive data analyzing method called file signature analysis is needed to support the process of Computer Forensics. This method is articulated in details in this article and discussed.

Introduction

Computer Forensics is the process of using scientific knowledge to collect, analyse and present data to courts. This process involves the preservation, identification, extraction and documentation of computer evidence stored in the form of magnetically, optically or electronically stored media. Steps in forensic process include:

1. Creating an exact physical copy of the digital media e.g. the computer hard disk. This is often called bitwise image
2. Load image to an empty or formatted hard disk
3. Secure the original media in a sealed container
4. Mark and retrieve data of evidential value
5. Present evidence in a readable form for court or tribunal

Step 4 involves the examination of the image and the search for evidence. With millions of files being stored on a computer, there is a need for methods to reduce the search space for the forensic examiners and spot out suspicious files. This is where signature analysis is used as part of the forensic process.

A signature analysis is a process where files, their headers and extensions are compared with a known database of file headers and extensions in an attempt to verify all files on the storage media and discover those which may be hidden. In order to fully understand the usefulness of signature analysis, this article gives an introduction to the structure of computer files and how such files can be hidden. Then, a demonstration would be articulated to show how signature analysis can be used to defeat such data hiding techniques.

Understanding the structure of a file

Since data are stored on computers as files, all of these files must be searched and examined as if they were files in an office for the purpose to gather digital evidence. In order to understand the process of

data hiding, one must first understand the structure of a computer file. The structure of a file normally consists of:

1. Filename
2. File header/footer
3. File content

1. Filename

The filename is a unique identifier which allows the computer to correctly identify each file stored on the disk. The first piece of information on the file format is given on the name of the file e.g. essay.doc. Different applications use different file formats to encode data on files so that other applications cannot extract the data. The part ".doc" is the filename extension. It is used by many modern operating systems such as Mac OS X and Windows to determine the format of the file and associate a list of application of which the file is compatible with.

2. File header/ footer

The information which describes the type of the file, e.g. which application the file is associated with, is stored in the header or footer (or both) within a file. Such information is called the signature of the file or file signature and they most often unique to one another. The file extension and the file signature of each file should match each other in most cases but there are a few exceptions. There can be mismatches, no match, unknown types and anomalous results.

Every file has a file header/footer which contains information on the format of the content stored in the file. It could either form a part of the file or stored as a separate file. Files of a particular type can be searched for using the information stored in the file header alone. Such information can be easily obtained by opening files using a hex editor such as HexEdit. Such tools allow users to see and edit the raw and exact contents stored in a file. Below is an Adobe PDF file opened using a hex editor:

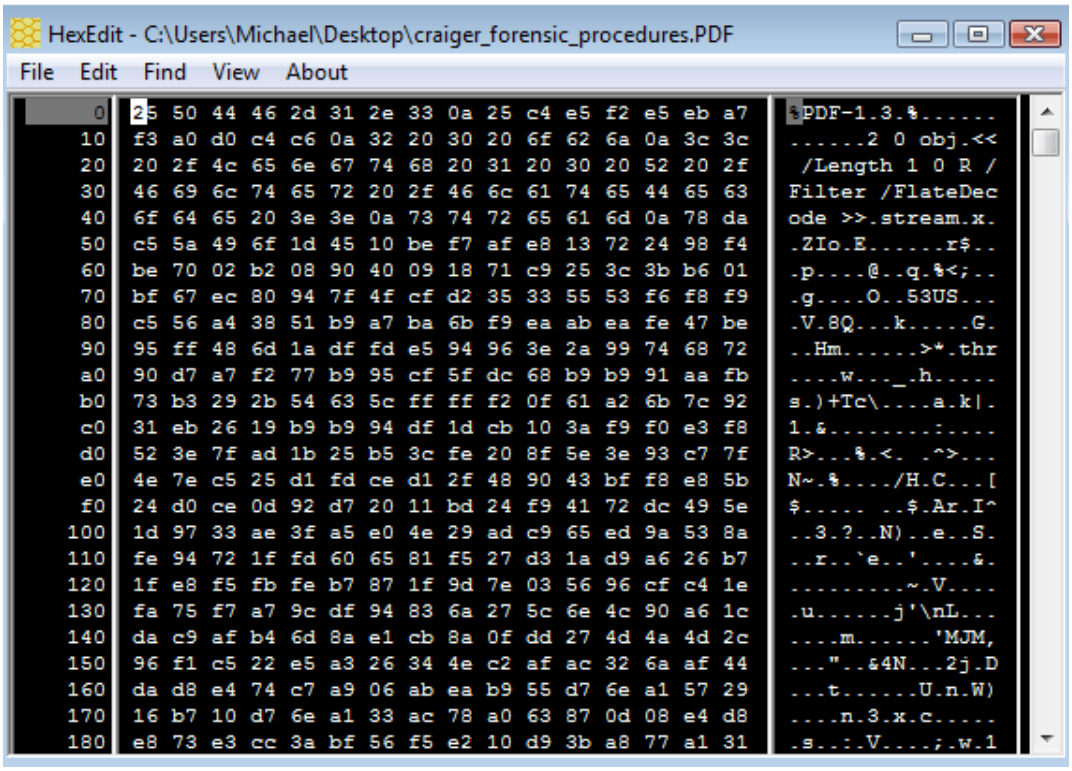

Fig 1. Hexadecimal representation of a Adobe PDF file

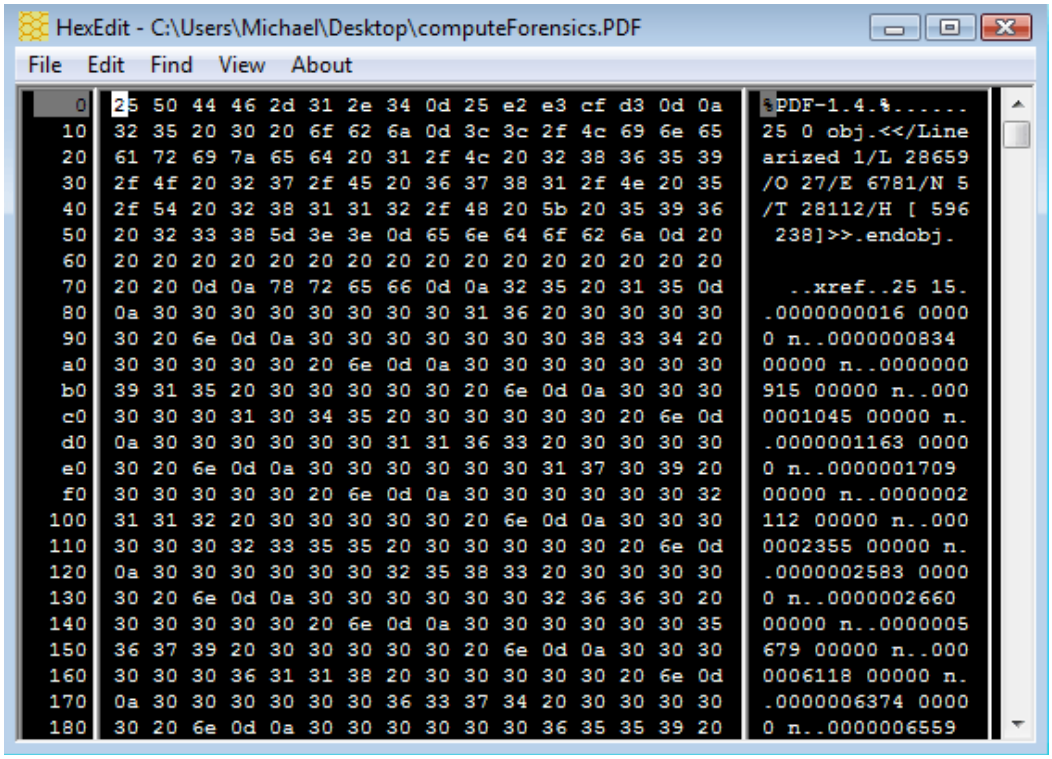

Fig 2. Hexadecimal representation of another Adobe PDF file

From the Fig 1., it can be seen that the header of the file is stored in the Adobe PDF file as displayed on the right column whilst its hexadecimal representation is displayed in the middle column. In order to find the signature of an Adobe PDF file, another Adobe PDF file was opened using HexEdit. Below shows the hexadecimal representation of another Adobe PDF file.

By comparing and contrasting Fig. 2. with Fig. 1., it can be deduced from that the first five hexadecimal values are the same for both documents. This is correct as this string of hexadecimal values contains the file signature of an Adobe PDF document. In fact, the official signature of a PDF document is the first four hexadecimal values, `25 50 44 46`.

Below is a list of publicised file signatures of some of the most common file formats:

File format	File signature (hexadecimal)
Adobe PDF	`25 50 44 46`
Microsoft Office 2007 Documents	`50 4B 03 04 14 00 06 00`
Microsoft Office 97-2003 Documents	`D0 CF 11 E0 A1 B1 1A E1`
JPEG images	`FF D8 FF`
GIF images	`47 49 46 38 37 61` ***or*** `47 49 46 38 39 61`

Table 1. List of file signatures of the most common file types

A file signature analysis makes use of a more extensive list of such file signatures to detect file tampering. This process is detailed in later sections.

3. File content

Since the purpose of the file is to store data for different applications, the major content of a file is the data from the associated application. Such data maybe encoded in various different ways so that only the eligible applications are compatible and competent to read and extract the content of the file.

Data hiding methods

The most common ways in which data can be hidden on computers are:

1. Hiding document inside another by changing filename extensions
2. Deleting the data

1. Changing filename extensions

Some information may be of such value that criminals may not want to delete them. Instead, they opt to hide them in such a way that the files may appear to other users as another type of document and only the criminal can retrieve the original document. Criminals can attempt to hide their confidential information by changing the extension of a file e.g. change an image file from .jpg to .doc or vice versa.

In Windows, one can easily merge a document with an image together using the command prompt using the command: `copy /secret mySon.jpg + confidential.pdf xmas08.jpg`. Windows would recognise the merged file e.g. `xmas08.jpg` as an image file and when double clicked, the image is displayed. The criminal can retrieve his document simply by changing the filename to the filename of his original document e.g. `confidential.pdf`. This trick would be enough to trick the eye in the Windows operating system since Windows would display the file as an image. Therefore, in order to search for such data, a more comprehensive search than the naked eye is required. Please note that in order to hide files with extensions DOC, PPT, WAV and other formats using this technique, compression into RAR format is required in order to protect the integrity of the original document.

2. Deleting data

This is by far the most common way of hiding data. Truly speaking, the user's original intention was probably to delete and destroy the data rather than hiding the data. However, deleted data are not "deleted" in a way that the data would be destroyed and irrecoverable at an instant.

So, how are files deleted? When a file is created, a directory entry for the file is also created. When that file is deleted (not deposited in a recycle bin like the one in Windows), the first character of the filename in the directory entry is changed to a special character (represented in hexadecimal as E5). Then, a search in the File Allocation Table for any entries with this filename is carried out and any entries found would be cleared. This process is simply a notification to the memory management unit that the memory allocated to this filename becomes available and can be reallocated to other processes. Until these memory slots are reallocated and overwritten by new data, the original data which was stored in the file would remain on disk and theoretically speaking, it can reside on the disk forever.

Procedures in file signature analysis

As seen from the previous section, files can be hidden and a search through a disk is not enough. One of the data hiding technique was to modify the filename extension to trick the operating system to believe that the file is of a different format. The purpose of the file signature analysis is to detect

whether a filename extension has been tampered with and once detected, a further investigation can be carried out on such files, narrowing the search space.

Normally, the file signature analysis is carried using forensic applications such as EnCase which enables the user to examine a disk image and carry out several different procedures. Such applications make use of an extensive list of publicised file signatures and match them with files' extensions. If a mismatch is found then the file's extension has obviously been altered and the file would warrant a closer examination.

In this case, the concept of file signature analysis is demonstrated by examining file signatures using HexEdit rather than EnCase.

Case study: Hiding a Microsoft Office 2007 document

It is demonstrated in this case study how easy it is to trick Windows Explorer to display wrong file type simply by altering the file's extension and how the examination of the file's signature using HexEdit defeats such method.

Platform: Windows Vista
Method of hiding: Changing file's extension
Forensic technique: File signature analysis using HexEdit

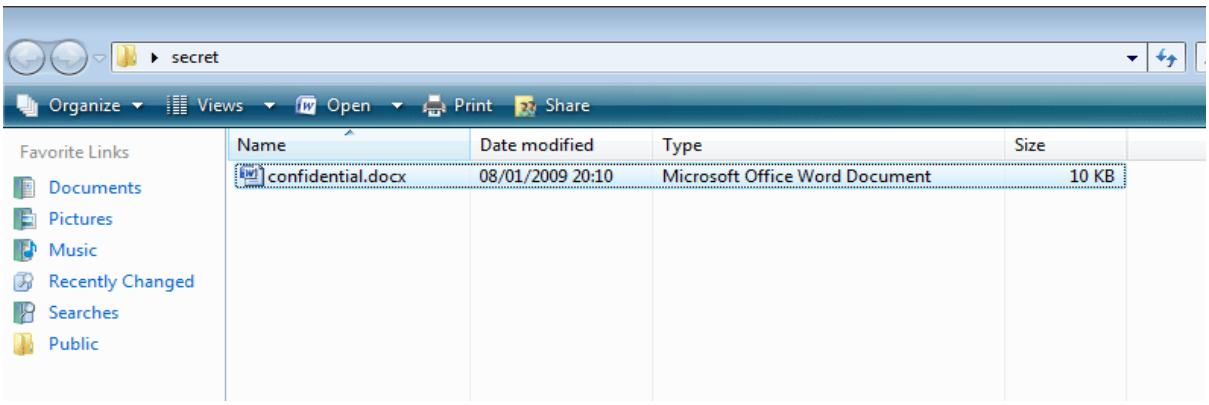

Fig 3. confidential.docx created in a directory named secret

Firstly, a new directory was created on the desktop called secret and a new Microsoft Office 2007 document was also created as shown in Fig.3. above. The document has a secret message "You shouldn't read this!" as shown in Fig.4. below.

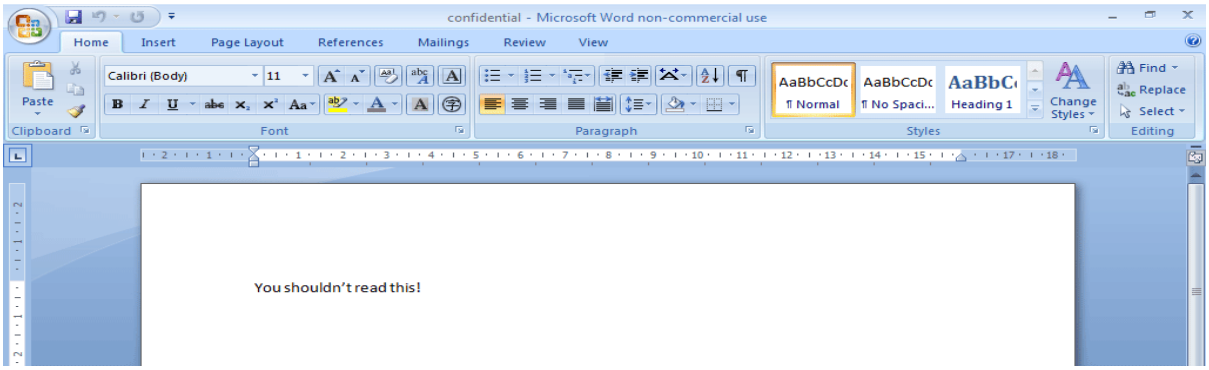

Fig 4. The message stored in confidential.docx

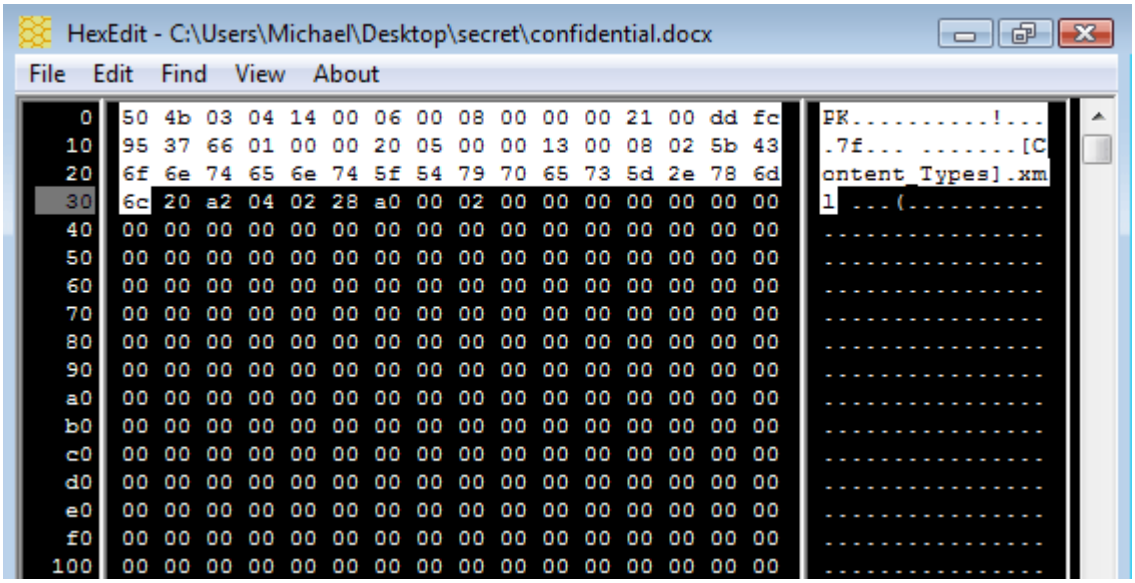

Fig 5. File header of confidential.docx

Examining the file header using HexEdit as shown in Fig.5.confirms the file signature of this Microsoft Office 2007 document is correct as listed in Table 1. in an earlier section.

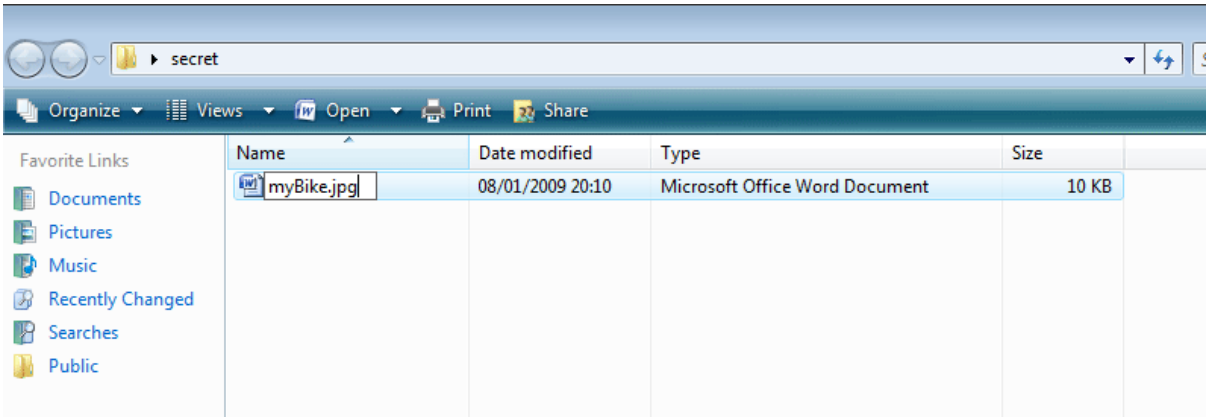

Fig 6. Renaming the file and changing the file's extension

Fig.6. shows the process of renaming the document from confidential to myBike and changing the file's extension from .docx to .jpg.

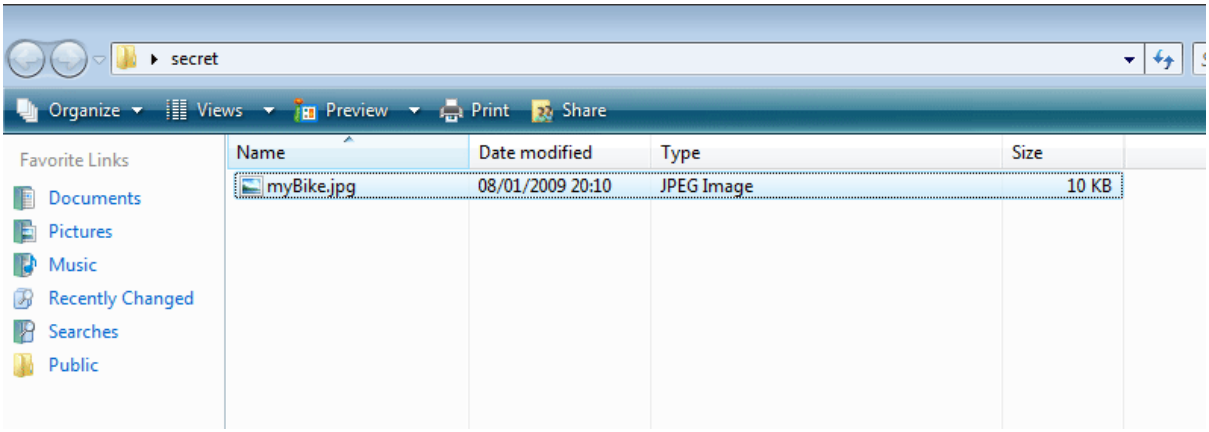

Fig 7. Windows Explorer is tricked to display file as JPEG image

Fig.7. shows that the renamed document is now displayed as a JPEG image rather than a Microsoft Office 2007 document. The Windows Explorer has been successfully fooled and examining the file with naked eye would not reveal the original document hidden.

Please note that data hiding using this technique does not affect the integrity of the original document as the `copy /directory` command is not used. However, when the image is double-clicked, no image is displayed and a warning message "The file appears to be damaged or corrupted".

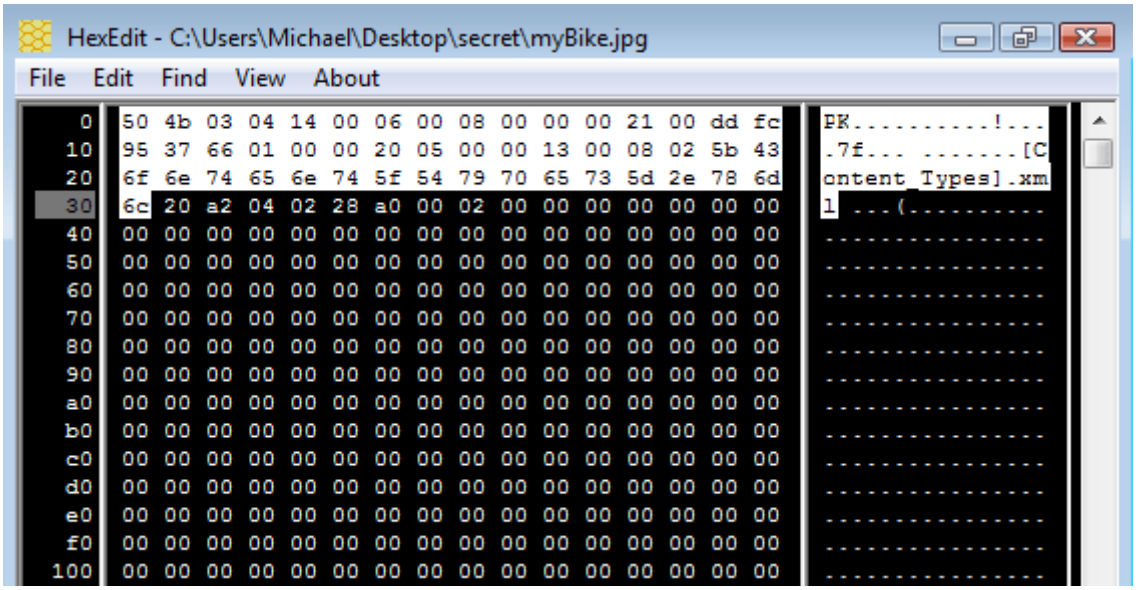

Fig 8. File signature of myBike.jpg belongs to a Microsoft Office 2007 document

When a forensic examiner sees this file, he would expect the file signature of this image is `FF D8 FF` since the file's extension is .jpg. However, by examining the file signature of the "image" file using HexEdit, the forensic examiner spots that the file signature and the file's extension has a mismatch and indicates that the file's change has been changed. This indicates that the file is suspicious.

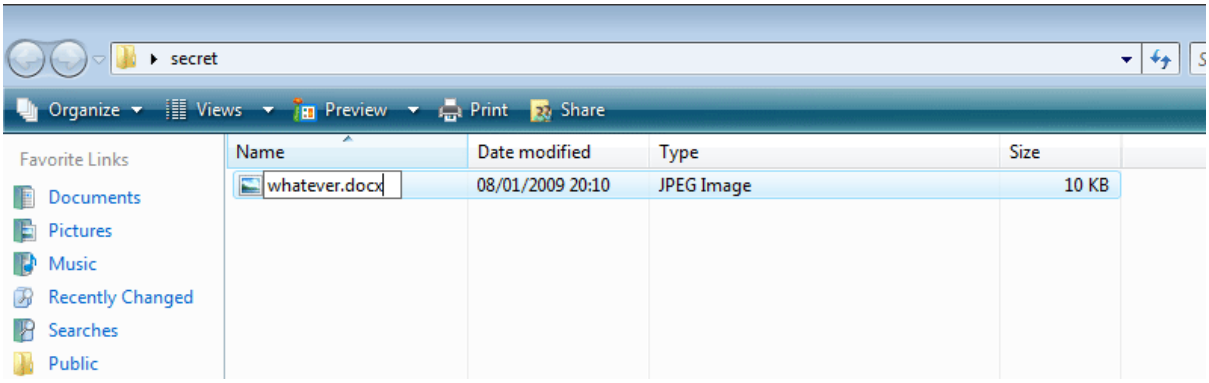

Fig 9. Attempt to recover original document by renaming and changing file's extension

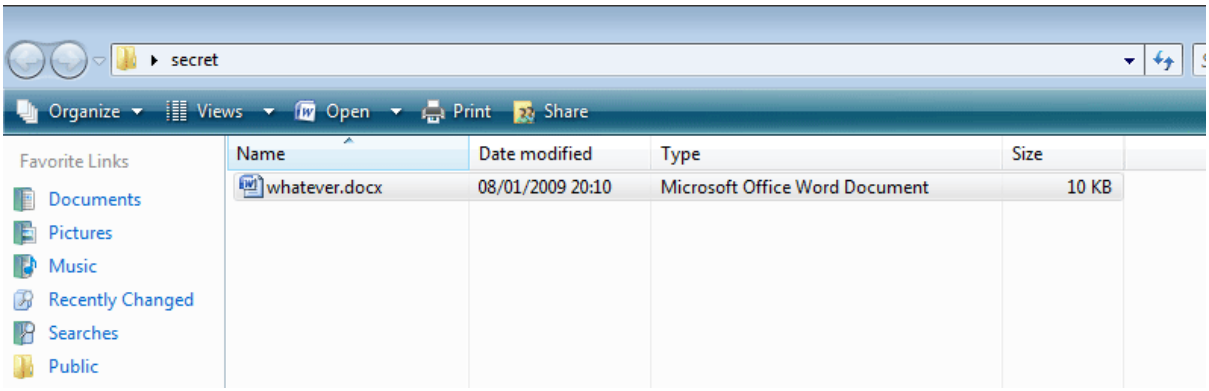

Fig 10. File has been changed from JPEG format to Microsoft Office 2007 format

The next step is to investigate what is hidden underneath the image. From the file signature found in the image file, it can be seen that the original document could in fact be a Microsoft Office 2007 document. The forensic examiner can simply rename the image file from myBike to whatever the examiner wants, e.g. whatever and changing the file's extension from .jpg to .docx. This is shown in Fig.9. and Fig.10. above.

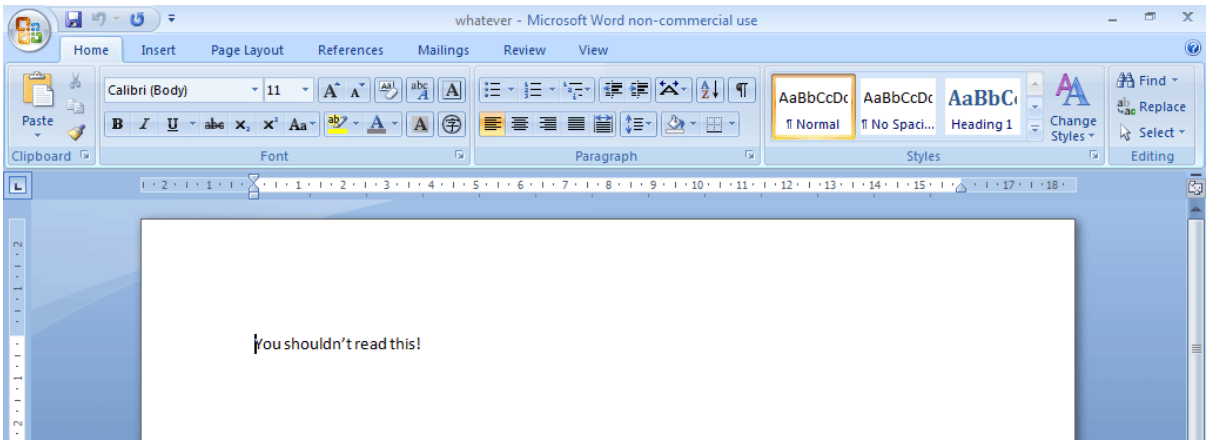

Fig 11. Secret message recovered

After changing the file's extension, the examiner opens the file with Microsoft Office Word and the secret message has been found.

This case study shows in detail the concept behind the file signature analysis. However, manually examining file signatures and matching files' extensions is too time consuming and this process can be fully automated using applications such as EnCase.

Other methods

File Signature Analysis is used for two purposes:

1. Spotting file signature and file extension mismatches
2. File carving – a process to identify file headers and footers using predefined file signatures. This is typically used to find deleted files.

However, with millions of files stored on the hard disk and some of which are system files, a process can be carried out before File Signature Analysis is carried. This process is called hash analysis.

Hash analysis works by comparing hashes of files from disk with a list of predefined file hashes. Those which match can be of two categories, known and notable. This process simply automates the process of finding those files which can be ignored e.g. typical system files and those which can be of evidentiary value e.g. Internet browser history files

Conclusion

It can be concluded that File Signature Analysis is a very useful technique in forensic computing. However, there is a big weakness with this process and it is that it relies entirely on the list of predefined file signatures being updated and contains every necessary file signatures. If this list is incomplete, valuable files maybe overlooked or ignored by applications such as EnCase and valuable evidence would be lost.

References

[1] Craiger, J *Computer Forensics Procedures and Methods*. Floria: University of Central Florida

[2] (2008.) *Computer Forensics.* Washington: US-CERT

[3] Pladna B. *Computer Forensics Procedures, Tools, and Digital Evidence Bags: What They Are and Who Should Use them.* East Carolina: East Carolina University

[4] Haggerty J. *Digital Fingerprinting For Computer Forensics*. Liverpool: Liverpool John Moores University

[5] Casey E. et al. (2001.) *Handbook of Computer Crime Investigation.* Academic Press

[6] Bunting S. (2007.) *EnCase Computer Forensics*. John Wiley & Sons

[7] Harris R. (2007.) *Using Artificial Neural Networks For Forensic File Type Identification.* West Lafayette: Purdue University

[8] Mohay G. et al. (2003.) *Computer Intrusion Forensics.* Artech House

[9] *HexEdit.* http://www.physics.ohio-state.edu/~prewett/hexedit/ [Accessed 08 Jan 2009].

[10] (2008.) *File signatures table*. http://www.garykessler.net/library/file_sigs.html [Accessed 08 Jan 2009].

Justification for the references used

[1] Craiger, J *Computer Forensics Procedures and Methods*. Floria: University of Central Florida

Computer Forensics Procedures and Methods is an article written by J. Craiger who is the Assistant Director for Digital Evidence. He is part of the National Center for Forensic Science and Department of Engineering Technology at the University of Central Florida. This article was used because it documents the procedures used in Computer Forensics in a very clear and easy-to-understand manner. It has a detailed section on signature analysis which enabled to fully understand the procedures in a file signature analysis and the role is played. Since the purpose of this article is to serve as a technical introduction to fundamental procedures in computer forensic, the content in this article is of a descriptive nature rather than opinionated. This led me to believe that the content is safe to use however, I have also used other sources to cross examine the content of this article.

[2] (2008.) *Computer Forensics.* Washington: US-CERT

This article was used to consolidate my understanding of the general principles of computer forensic. Since this article was produced by US-CERT which is a US government company, its content is safe to use for this essay.

[3] Pladna B. *Computer Forensics Procedures, Tools, and Digital Evidence Bags: What They Are and Who Should Use them.* East Carolina: East Carolina University

This paper was written by Brett Pladna who was a Graduate Assistant at East Carolina University. This article is of a descriptive nature and hence contains very little opinions. It was used to cross examine other articles.

[4] Haggerty J. *Digital Fingerprinting For Computer Forensics*. Liverpool: Liverpool John Moores University

This is a set of presentation slides on the subject of computer forensics. It was used to point me to the correct directions in my research for more information on signature analysis and the overall discipline of computer forensics.

[5] Casey E. et al. (2001.) *Handbook of Computer Crime Investigation.* Academic Press

This book was written by Eoghan Casey who has extensive experience in the computer forensic field and it was used to give me a bigger picture on the discipline of forensic computing. It also introduced to me the concept of signature analysis and hash analysis. This book was used to cross examine the materials in the *Computer Forensics Procedures and Methods* article.

[6] Bunting S. (2007.) *EnCase Computer Forensics*. John Wiley & Sons

This is a book written specifically for people who wish to take the EnCE Certification, the official EnCase examiner. This suggested to me that the material in this book is the official material on the subject of computer forensics. Also, it has an extensive chapter on signature analysis and shows how EnCase is used to perform such process. Since this is the official publication for EnCE Certification, it was used to cross examine all other materials I have used in this essay.

[7] Harris R. (2007.) *Using Artificial Neural Networks For Forensic File Type Identification.* West Lafayette: Purdue University

This article was written by Ryan Harris who is Systems Engineer II at Verizon. This article was written for CERIAS, the Center for Education and Research in Information Assurance and Security, Purdue University. This article highlighted to be some of the vulnerabilities of signature analysis. As this was a report from his work, I felt that there was a need for some cross examination and so I used

[8] Mohay G. et al. (2003.) *Computer Intrusion Forensics.* Artech House which was written by George Mohay et al. in an attempt to provide a detailed introduction to the discipline of computer forensics. I used this book to cross examine the vulnerabilities mentioned in Ryan Harris's article.

[9] *HexEdit.* http://www.physics.ohio-state.edu/~prewett/hexedit/ [Accessed 08 Jan 2009].

HexEdit is one of a few hexadecimal file editors available on the internet. It was used to examine file signatures.

[10] (2008.) *File signatures table.* http://www.garykessler.net/library/file_sigs.html [Accessed 08 Jan 2009].

This site contains a comprehensive list of file signatures. I used this to find the file signatures of some of the most common file formats. I cross examined file signatures found on this site by opening some files of particular formats using the HexEdit application.