MANAGING LARGE AMOUNTS OF ELECTRONIC EVIDENCE

*Cybercrime Lab**
Computer Crime and Intellectual Property Section
Criminal Division

Ovie L. Carroll, Director, CCIPS Cybercrime Lab
Stephen K. Brannon, Cybercrime Analyst, CCIPS Cybercrime Lab
Thomas Song, Senior Cybercrime Analyst, CCIPS Cybercrime Lab
Joel M. Schwarz, Trial Attorney, CCIPS

I. INTRODUCTION

Investigations usually focus on finding and getting evidence. A computer-related investigation often generates a particularly large amount of evidence. Managing all this data and using it effectively through the lifecycle of an investigation presents special problems. This article explores those problems and describes general strategies and some specific solutions for managing large amounts of electronic evidence.

The Cybercrime Lab in the Computer Crime and Intellectual Property Section regularly provides advice and assistance with the issues in this article. You can contact the Lab at (202) 514-1026 or at www.cybercrime.gov. Also, from within the Criminal Division or US Attorneys' offices, you can access resources on our intranet site, CCIPS Online, by going to DOJ Net and clicking the "CCIPS Online" link.

II. CONCEPTS AND CONCERNS

A. Preliminary Concerns

There is one cardinal rule for electronic evidence: always work on a copy. Original evidence or the single best copy should be duplicated and kept safe. A clean chain of custody record should be maintained for that best copy. Only use working copies of evidence for review and analysis. Always keep the original or best copy safe and disturb it as little as possible.

Working directly with original evidence or your best copy is extremely dangerous. The first reason is that simply interacting with it will likely change it. It is also dangerous because there is a greater risk that data will become corrupted or lost due to hard disk failure. The integrity of electronic evidence is important just as with other types of evidence. But the integrity of electronic evidence is also important because if it is intact, forensic copies should theoretically be exact. With some other types of forensic evidence, testing and analysis use up the evidence itself. But with electronic evidence, any number of exact copies can be made, and the defence is often entitled to receive a copy for review.

If the government can't produce an exact copy of the evidence that it seized or obtained for any reason, it opens a Pandora's Box of questions about what went wrong. Even if someone accidentally modifies evidence, it is still likely admissible. If the modification is clearly documented and explained then the evidence can probably be used. However, the modification may influence the evidence's weight.

Electronic evidence is much easier to manage if a system of organization is already in place before it is collected. As investigations get evidence, they naturally document and preserve original versions. But as copies are made, it is helpful to plan a system of organization and start filing copies of evidence into it. Far too often, investigations let the order in which they get evidence or its sources dictate its organization. Then at the end of the investigation when the evidence needs to be sorted differently to be used, there may not be time to reorganize it.

---

* The Cybercrime Lab is a group of technologists in the Computer Crime and Intellectual Property Section (CCIPS) of the Department of Justice, in Washington, DC. The lab serves CCIPS attorneys, Computer Hacking and Intellectual Property (CHIP) units in the US Attorneys' offices, and Assistant US Attorneys in general, by providing technical and investigative consultations, assisting with computer forensic analysis, teaching, and conducting technical research in support of DOJ initiatives.

The idea is to think forward to your analysis so it can guide your initial setup. For example, you may be starting an investigation of multiple targets using multiple websites. If you know that most of your questions will be about one target or another, then you would organize evidence by target as you get it. On the other hand, if you know you will need to paint a coherent picture of the activity on each website, you would organize evidence by website. Planning and setting up an organizational system at the beginning of an investigation may determine whether or not electronic evidence is manageable at the end.

Another preliminary concern for electronic evidence is the use of date and time information. Problems with computer date and time settings can be fatal to an investigation. Overlooked date and time issues can ruin an investigation quickly. Targets can be misidentified, evidence can show a target did something he or she did not really do, and evidence from different computers can be inconsistent. They are all too common, but they are still often overlooked. Every computer, server, etc. has an internal clock. The date and time - or at least what the computer believes they are - will be spread through all the evidence the computer produces, especially any logs. The clock setting may be set wrong or it may be set to a different time zone. Investigations need to find and adjust both for inaccuracies of any particular clock and for discrepancies between different clocks.

Fortunately, it is possible to document and compensate for almost any problem with dates and times as long as it's both *identified* and *quantified*. Say you are running an undercover website and logging the activity that takes place on it. You may discover that the computer running the website had an incorrectly set clock and it was set exactly 23 minutes fast for the last year. With both of those pieces of information, the year's worth of log evidence can be salvaged and used by subtracting 23 minutes from every time.

A final idea that should guide everything you do with electronic evidence is: *let the computer do the work*. You may find yourself in a situation where you are tempted to have an army of investigators or paralegals process mountains of evidence manually. This is almost always the wrong answer. It's far better to think of a way for the computer to do brute force searching, sorting, etc.

A human brute-force attack is too slow, but it also introduces the potential for too many errors. Even the most conscientious person can't avoid making mistakes when he or she has to do the same thing 1,000 times. On the other hand, once you give a computer accurate instructions, it can easily execute them a million times without any mistakes. A long list of answers without mistakes is often what you need from electronic evidence.

What are some things an investigation might need to do with electronic evidence? Thinking about this question helps structure our discussion. The first task is to organize and manage it. The next task is to search it. It often feels like trying to find a needle in a haystack. For example, an investigation may need search through millions of lines of logs to find the one record that shows a target did something on a website.

The third and final type of task is analysing and interpreting evidence. This entails looking at all of the evidence, or a large part of it, and synthesizing new information from it. This type of task is anticipated less often but can still provide the most useful information. For example, think again about an investigation of many targets using many websites. It might be necessary to identify the most egregious users to select targets for prosecution. We would need to quantify each target's conduct across all the websites throughout the course of the investigation and then compare the targets.

B. Indexing

Indexing is a technique to search through large bodies of data faster. Indexing goes through the entire body of data and creates a map of what information is where. This map, or index, functions like the index in a book or the card catalogue in a library. Building an index can take a long time. But once it's done, searches can be done much faster. For example, Google and LexisNexis have indexed all their data to enable users to search it quickly. It is hard to imagine how long it would take to search every word on the Internet or every word in the LexisNexis databases if they had not already been indexed. In situations with a large amount of data where multiple searches will be necessary, it's generally best to index once and then use that index to search. In the long run, this is far faster than conducting each search through all the data.

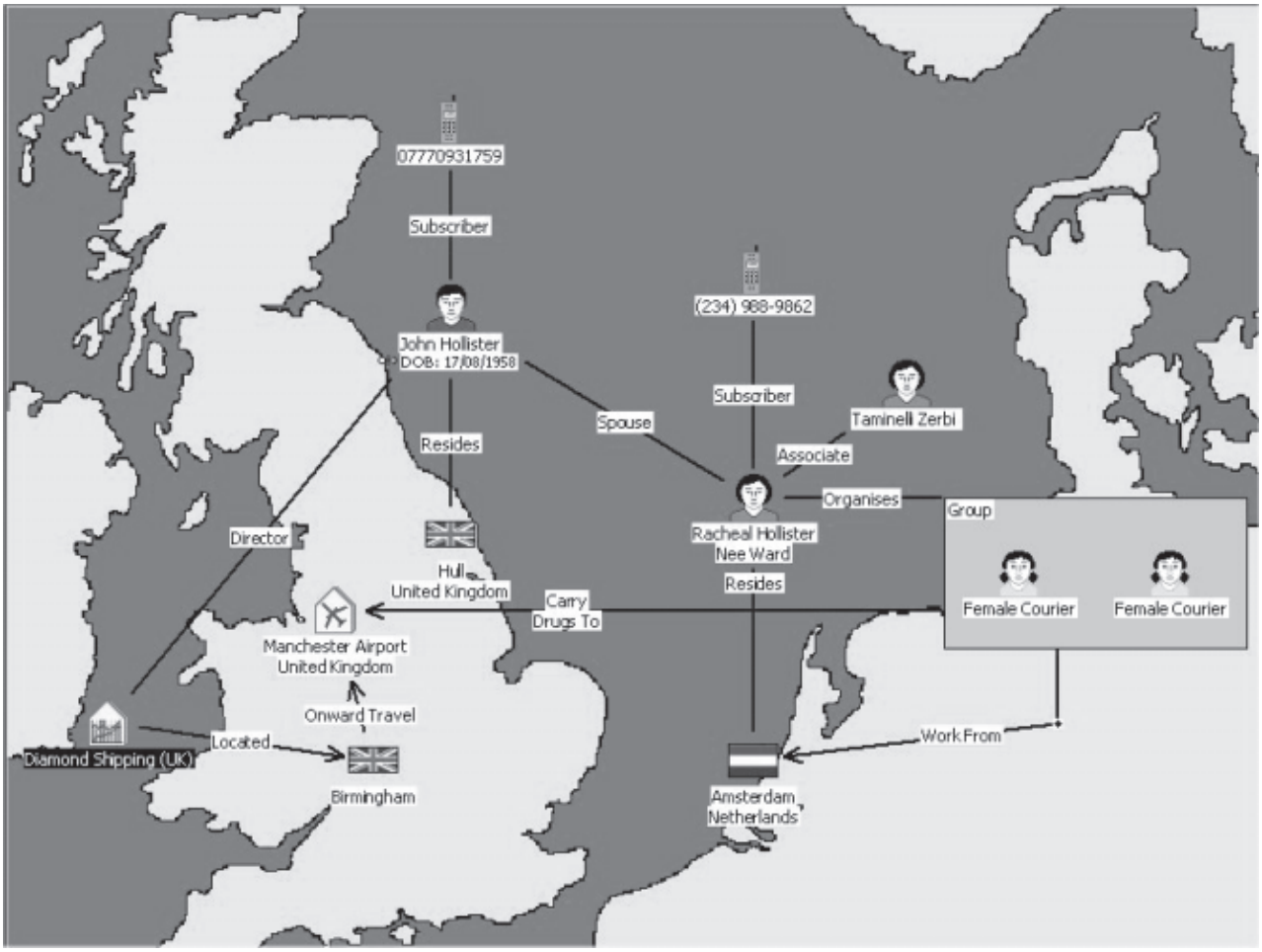
Indexed searching can be done within computer forensics programmes. It can also be done using stand-alone programmes that only index and search data. The computer forensics programme Forensic Toolkit

(FTK) made by AccessData has the capability to index data built in and is widely considered the leader in indexed searching. The other leading computer forensics software is EnCase, made by Guidance Software. Its newest version, version 6, also incorporates indexing capabilities. Indexing can be done in previous versions of EnCase by using a third-party add-on, such as Mercury by MicroForensics. Once data in a forensics programme has been indexed, searches that would have taken minutes or hours are completed almost instantly.

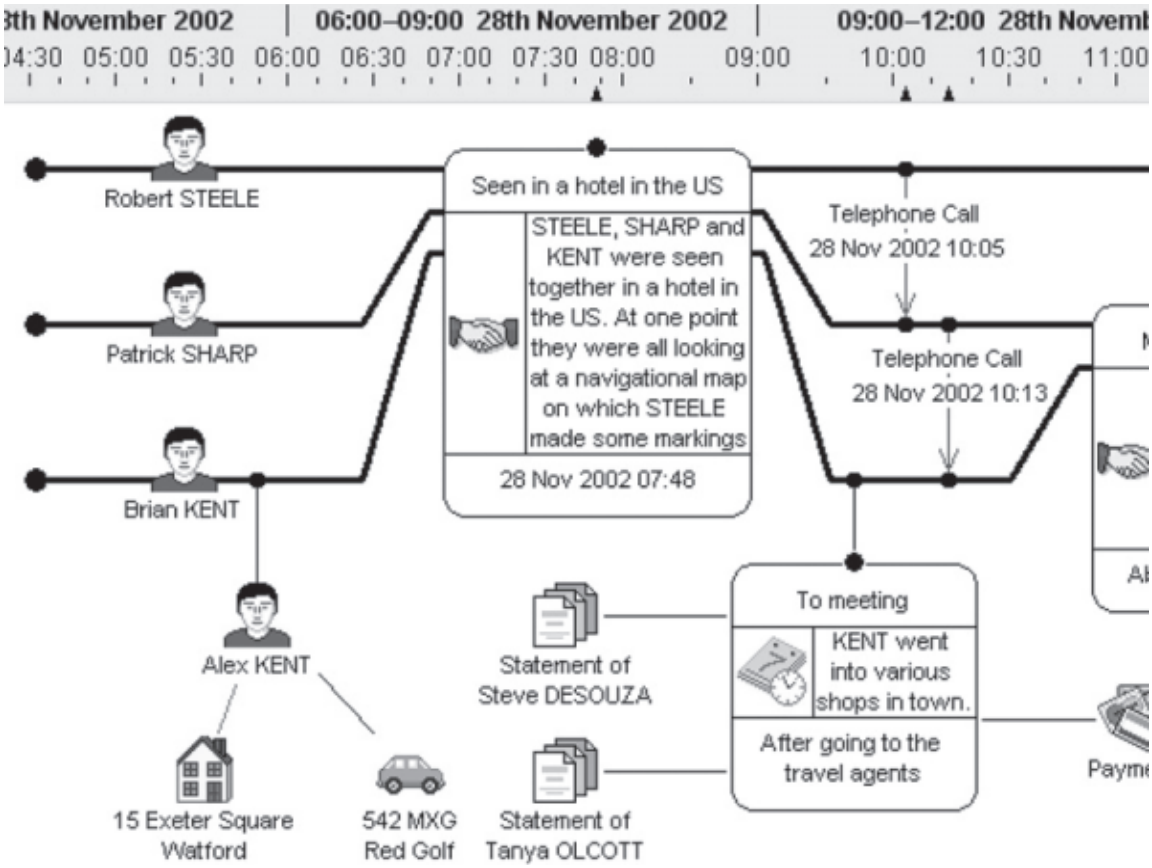
There are also stand-alone programmes that just do indexed searching. DtSearch produces a mature suite of programmes that use the same indexing engine as FTK. The basic programme searches text in multiple formats and highlights results. It also has options to use fuzzy, phonic, wildcard, stemming, and thesaurus search options. These are all search techniques that find results similar to or related to the term you provide. This can find misspelled occurrences of words, so it is especially useful when searching through anything written by a person. For example, a fuzzy search for "apple" would also find "appple". DtSearch can also display results as web pages that are easy to use. Another programme in the product suite, dtSearch Publish, lets you publish and distribute your evidence in an indexed and quickly searchable package. This product is excellent for giving indexed copies of your evidence to others to review.

C. Visualization

This article discusses many ways to manage, search, and analyse electronic evidence. Sometimes the only way to see important relationships is to visualize large amounts of information. Also, some results are only useful to a prosecutor or jury when they are presented visually. There are programmes available that combine database and visualization features to enable an analyst to find connections and illustrate them. These tools are often used in cases with extensive financial data or phone records. They are also particularly useful to show relationships indicated by email exchanges or network traffic. One of the most popular programmes is Analyst's Notebook by i2 (http://www.i2.co.uk/Products/Analysts_Notebook/default.asp). It can illustrate relationships as shown in the screen capture below.



The analyst's notebook can also perform and illustrate timeline analysis. An example of this is below.



III. TECHNIQUES AND TOOLS

A. Email

It is important that electronic evidence only be reviewed on a computer that's not connected to the Internet. It is most important when reviewing email evidence. Reviewing email on a computer connected to the Internet risks accidentally sending a read receipt response to addressees on the email. Also, some email uses HTML, the language for making web pages, to control formatting. In fact, Outlook created message in this format by default. Since HTML can have references to images and other files on websites, simply opening it can cause your computer to connect to those websites to retrieve message elements. This can directly or indirectly warn a tech-savvy target that he or she is under investigation.

Reviewing email on a computer connected to the Internet even risks sending an email to a target tipping him or her off. We know of a case where the agent reviewing an email between conspirators accidentally double-clicked the "reply-all" button. Worse yet, the agent was reviewing the email in his own email account on his work computer. So he created and sent an email from his work email address to all the conspirators jeopardizing the investigation. In fact, a computer used to review email should not only be disconnected from the Internet, but it should also be dedicated to offline evidence review. An offline computer may keep track of read receipts that it is unable to send, then if it is later connected to the internet, it will take the opportunity to send them all.

Email provides several sources for valuable information. Some will often be in the body of an email. In addition to content, one can search or sort email by elements of the header, like the sender, recipient, subject, or date sent. It may also be helpful to search or sort by other less obvious attributes like the number of attachments, attachment names, priority, or age.

Once an investigation gets email from one or more sources, the first step is usually to import it all into one email programme. This facilitates organization and management. Sometime an investigator will instead review email one message at a time or import different groups of email into different email programmes. But this makes managing the evidence harder and searching it harder still. When all email is in one programme, one can easily organize it by folders in a structure that makes sense. Then one can conduct searches across all email or just across certain folders.

There are many free and commercial email programmes available. In our lab's experience Mozilla Thunderbird is one of the best free programmes for managing and searching email for all but the largest cases. The programme is free and available at http://www.mozilla.com/en-US/thunderbird/. Other programmes that are either free or likely already installed on most computers include Outlook, Outlook Express, and Eudora. The largest cases may need to use specialized forensics programmes like Access Data's Forensic Toolkit and Paraben's E-mail Examiner.

The rest of this section about email will describe the steps to import, manage, and search email. A preliminary step is needed for most email programmes. In order to operate, most programmes need the user to create a profile. This simply involves entering a name, email address, and a few other pieces of information. When a programme first starts, it usually walks the user through the account creation process. Made-up information is fine for this since the computer won't be connected to the Internet anyway.

Email in the most common formats can be imported into Mozilla Thunderbird. One common format is the mbox format. Such files are easily recognizable by the file extensions .mbx or .mbox. In fact, the mbox format is very common, and if a file with email has no file extension, it is likely an mbox file.

Thunderbird uses the mbox format internally, so the simplest way to import emails in that format is to copy the file to the directory where Thunderbird stores its own files. Then the next time the programme starts, the mbox file and all its email appears as a folder under "Local Folders".

On Windows XP, the directory is C:\Documents and Settings\[User Name]\Application Data\Thunderbird\Profiles\xxxxxxxx.default\Mail\Local Folders\ (xxxxxxxx is 8 random characters).

On Windows Vista, the directory is C:\users\[User Name]\AppData\Roaming\Thunderbird\Profiles\xxxxxxxx.default\Mail\Local Folders\ (xxxxxxxx is 8 random characters).

Just copy email evidence files to that directory, then open the programme and it's ready to use.

Another common email format is Micrsoft's .pst (Personal Folders) format. Thunderbird can't import a .pst file directly, but it can be imported into Microsoft Outlook first and then from Outlook into Thunderbird. Again, it is essential to first create a profile in each programme as described above. Importing a .pst file into Outlook only takes a few steps. The goal for this step is to get email from a .pst file into the Personal Folders in the Outlook profile. These instructions are specifically for Outlook 2003.

1. In Outlook, click the File menu then Data File Management.
2. Click the Add button then click OK.
3. Find and select the .pst file.
4. Optionally, type in a new name in the "Name" box (for example, "warrant response") then click OK.
5. Click Close.

Now the .pst file should appear as a folder on the bottom of the left pane (in our example, "warrant response"). The last thing to do in Outlook is move the mail from the new folder to a folder inside the Personal Folders. To do that:

1. Right-click on Personal Folders and select New Folder.
2. Name the folder (for example, "Bad Guy1") and click OK.
3. Click on the .pst folder at the bottom ("warrant response").
4. Select all the emails by clicking the Edit menu then selecting Select All.

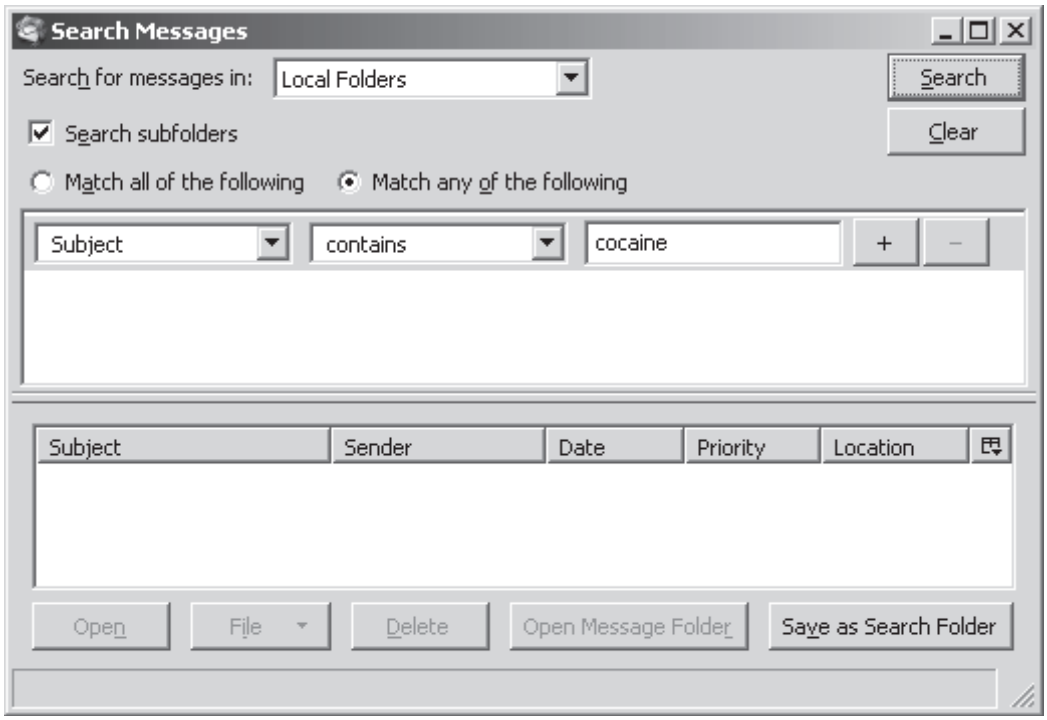
5. Carefully click on any email and drag it into the folder created in the Personal Folders ("Bad Guy1"). This should move all email from the .pst file into the local folder.

Now you can close Outlook and open Thunderbird to import the email from Outlook. These instructions are for Thunderbird 2.0.0.6.

1. In Thunderbird, click the Tools menu then Import.
2. Select Mail then click Next.
3. Select Outlook then click Next.
4. The process imports every folders from Outlook, so it may be helpful to delete empty or unrelated folders.

Bringing all email evidence into one programme has two main advantages. First, you can organize and manage it in a way that works best for your case regardless of how you got it. Second, and even more beneficial, you can search through all email evidence at the same time or search only through sections that make sense.

To open Thunderbird's search interface click Edit, then Find, then Search Messages. The search interface looks like this:



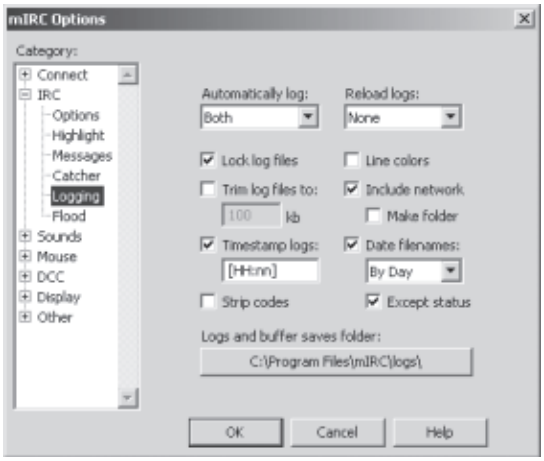
The box at the top selects which folder or folders to search. Select "Local Folders" and leave the "Search Subfolders" box checked to search in all email. The two radio buttons and the middle pane specify search conditions. The radio buttons determine whether *all* the conditions must be met for a result to be included or if it will be included when *any* of the conditions are met. Each search condition specifies where in the email to look, the condition to meet (i.e., contains, doesn't contain, begins with), and the search term. You can easily add or remove any number of conditions by clicking the + or – buttons.

Click the Search button and search results are displayed in a list in the bottom pane. This list can be sorted by any field. A search hit is easy to file into a folder. For example, you can create a folder called "key emails." Then when you select an email in the search results list, you can click the File button on the bottom and select the folder you want to move it to. You can also save a useful search. Click the Save as Search Folder button and it will create a search results folder. The results will be viewable as if they were a folder, but the original email won't be moved.

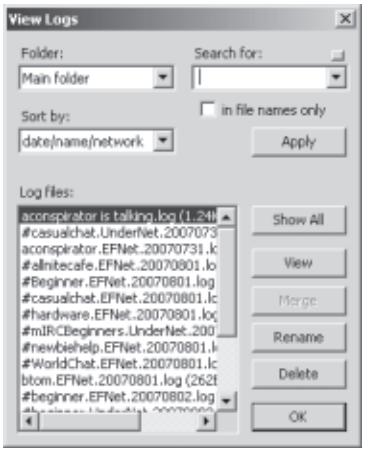
B. Chat Logs

Many computer-related investigations involve records of online chat, or instant messaging. Instant messaging lets two or more people to have a real-time, text-based conversation over the Internet. Each user types messages into a window on his or her computer, and every user who is party to the conversation sees all the typed messages in real time. So for example in a one-on-one chat, two people send text message back and forth. Both people see the conversation scroll by in a window. Or in a chat room, or channel, with several people communicating, each person sees a window representing the room and everything everybody types is visible. Most instant messaging programmes allow users to log their chats and some programmes even log by default. You may get chat logs from a target's computer, from a victim, or you may record them yourself using a co-operator or undercover agent.

The programme mIRC is the most common programme used for IRC chat. IRC is a type of chat popular in many tech-savvy crime circles, such as hacking, ID theft, and high-level copyright infringement. The mIRC programme conveniently has an option to log its communications. The person using the programme only needs to check the right boxes and the programme produces its own logs and organizes them in folders. So even as it creates the logs mIRC can already introduce a level of organization.



The programme also has an interface for viewing and searching its logs. Even if you later reorganize the chat logs into a different folder structure that is easier to manage, this interface can still see them and interact with them in the same way. It looks like this:



Extensive chat log evidence may require powerful techniques like those described in the next section of this article. But for smaller collections of chat logs, this interface allows you to perform basic searching,

sorting, and analysis. The controls at the top search and filter which chat logs are listed. The bottom of the interface lists log files that meet the criteria in the top half. It provides ways to view and manage them. You can open any chat by double-clicking on it. By default it will open in a text editor, like Notepad, where you can again search for specific terms within that chat. It is also possible to merge related logs into a single file. You can select multiple files from the list (or all of them), and then click the Merge button. This combines the selected files into one new file.

Here are several examples of how you can use this interface to search and analyse chat logs. If you want to see who you have logs of conversations with, you can simply sort by name. If you want to see what a particular target has been chatting about, you could search for his or her user name. The resulting list of log files would be the chats in which he or she spoke. Then you could merge these into one file and send it to someone else for further analysis. You can find out who was talking about a particular topic and when they were doing it. You could do this by searching for a term linked to the topic. Sorting the results by name first and date second would show you who was involved in conversations about the topic and when the conversations took place.

C. Logs

Sometimes commercial off-the-shelf programmes are best for managing electronic evidence of the type they are designed to manipulate. Sometimes a programme can even manage its own logs. Often simple solutions like this are best. But some types of evidence have no readily available programme to manage them. Or if a programme is available, it may not do everything needed. This is often the case with raw log evidence. Log evidence is a file generated by a computer that records events, usually sequentially. These files can be logs of system events, such as every time a user logged on. They can also be logs of activities, for example, a file server may log every file transfer. Networking elements like firewalls can also generate logs that record activity on a network. These log files can easily be millions of records long or longer, and normal tools and techniques for managing them quickly become insufficient.

Microsoft Excel can open smaller log files, but it has several limits to its usefulness. First, in versions up to Excel 2003, a worksheet could not have more than 65,536 rows. Many log files have more lines. Excel 2007 can now have up to 1,048,576 rows, so it can at least theoretically open most typical log files. A second limitation is that when Excel opens a file it attempts to load the entire file's data into memory at the same time. For large files this can be impractically slow. Excel's final limitation is that its search and analysis capabilities are far inferior to those of databases.

The Cybercrime Lab has had great success using custom Microsoft Access Databases to manage log evidence. Using Access has several benefits: as part of Microsoft Office, it is already installed on most computers. It is reasonably easy for people with other technical experience to learn and use. Finally, it is powerful enough to handle all but the most voluminous log evidence (we almost never need to move to a more robust database with a bigger capacity). Not everyone is comfortable working with databases. But it is likely you can find someone in your organization with the aptitude for basic database work who can assist your investigation.

The same tools and techniques can be used for any kind of log evidence, but for the sake of clarity we'll discuss one type of log as an example. In our section we manage log evidence for many "warez", or online piracy cases. Targets in these cases often use file servers where each file transfer is logged. Each time a file is transferred, a line is written to a log file with information such as the date and time, the file name, the direction (upload or download), and the user's name. These log files can easily grow to be millions of lines long. Managing them as text files quickly becomes difficult and searching them or making sense of them quickly becomes impossible.

The logs are easy to manage once they are imported into a table in an Access database. The essential step is to split each line of the log into pieces and put each piece into a separate column in the table. So in our transfer log databases, a row in the table represents one line of the log. And every row has a column for each piece of information in it. For example there is a date/time column, a filename column, a direction column, etc. Splitting each log line into its parts is essential: it allows you to use the full power of a database. Depending on the format of a log file, Access may be able to import it and split each line into separate fields at the same time. Otherwise a simple Visual Basic module can parse the log files into pieces and perform any additional logic necessary while it imports them. Here is sample code for importing lines of a log file into a table in Access:

```
Public Function import(path As String)
    Dim rs As Object 'destination table
    Set rs = CurrentDb.OpenRecordset("tablename")

    Dim pcs() As String 'pieces
    Dim inp As String 'line read from input file

    Open path For Input As #1 'open file for input

    'import file
    Do While Not EOF(1) 'check for end of file
        Line Input #1, inp 'read line of data
        If inp <> "" Then
            'split line
            pcs = Split(inp)

            'put in record
            With rs
                .AddNew
                .field1 = pcs(0)
                .field2 = pcs(1)
                .field3 = pcs(2)
                'etc.
            End With
        End If
    Loop

    Close #1 'close file
    rs.Close
End Function
```

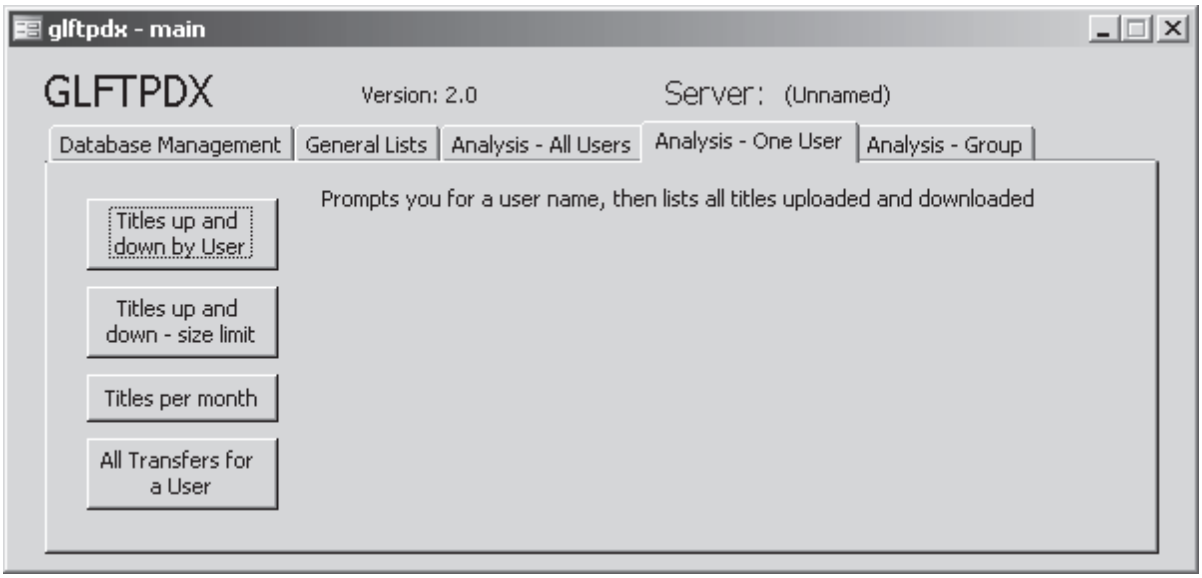
Once the logs are in a table in a database, you can do any searching or sorting by creating queries. A query is a structured way of getting data out of a database to answer a question. Access has a friendly interface to guide you through the process of setting up a query. You can select which fields you want in your answer, which fields you want sorted, and you can add any conditions. Keep in mind that Access can save any query you create and run it again at any time. This way, if you add or change data, you can ask the database the same question again and get the updated answer.

For example, sometimes we want to know who was using a particular file server. We made a query that told the database, “show just the user column, sort it in alphabetical order, and don’t show duplicates”. When we run the query, the database quickly generates a new set of data, like a mini-table, that answers the exact question we described. It gives us an alphabetical list of unique user names. Databases do this efficiently, so it easily runs through millions of records and gives us an answer in a few seconds.

We also often want to know what a particular target has done. We want a list of his transfers. To do this we made a query selecting three columns: file transferred, date, and user. We sorted the query by the date and limited it to one user (a condition for the user column). Again, we had another mini data set to answer our question in a matter of seconds.

Finally, sometimes we need to run a more complex kind of query. For example, we want to know who the most active users are on a server. In other words, who uploaded and downloaded the most. Our answer was a table with these three columns: user, count of his of uploads, and count of his downloads. Counting something for each user requires something called a crosstab query. Fortunately, Microsoft knows it’s a little more complicated so they provide a special wizard that walks you through creating one. You don’t even have to understand how it works: you just use the wizard to describe what you want.

In our lab, we had many related cases like this and many people needed to use the evidence. So we programmed a user interface to make the functions described above look like a friendly programme. A screenshot of the programme's main window is below. It uses tabs to group tasks (Database Management) and questions (General Lists, Analysis – All Users, Analysis – One user, etc.). On each tab there is a buttons for each query. If you point the mouse at a button it shows a brief description of what the query does. A database application with a user interface like this certainly isn't necessary for every case. But it may be appropriate when you need to harness the power of a database and make it available to a large number of non-technical users.



IV. CONCLUSION

We hope that the strategies and examples in this article will help prepare you to manage electronic evidence in your cases. The Computer Crime and Intellectual Property Section and the Cybercrime lab is also available to AUSA's for consultation on computer forensic and other technical investigative matters by calling (202) 514-1026. Many other resources are available on our section's public website, www.cybercrime.gov. In addition, anyone in the Criminal Division or US Attorneys' Offices can find additional resources on our new intranet site, CCIPS Online. Just go to DOJ Net and click on the "CCIPS Online" link. We also encourage AUSAs to take advantage of the many courses we present at the National Advocacy Center throughout the year.