Explanation of selected terms commonly encountered in computer-forensic discussions

John Butler, Geode Forensics Ltd.
3 December 2007

Contents

Internet Browser Cache and cookies

When a user browses the Internet with a browser program such as Internet Explorer the result seen is a set of pages laid out on the screen. Actually each page is usually made up of several, possibly many components – the text plus layout information is downloaded first as an *HTML* page then files containing each button, image, graphic, piece of decoration and so on is downloaded in turn and displayed in its proper location. The browser automatically makes a local copy of each of these components on the hard disk so if a similar page is downloaded or the same page requested again the components can be retrieved from this hard disk *cache* rather than the (generally much slower) network.

The system periodically cleans out old material to make way for newer items and maintains an index of what is in the cache. From a forensic point of view the cache can give a very good idea of what the user was browsing and depending on usage pattern can give a partial history going back months or years. The cache is an artefact for speeding up browsing not an audit trail so it is incomplete but can still give a good idea of what was going on.

Filenames in the internet history also show artefacts such as numbers in square or round brackets and replacement of spaces by “%20”.

The cache mostly records the components of pages being brought back to the user but can also record the *search terms* entered into search engines such as Google as these will be formed into requests for material. These will generally start with the URL (network location) of the search engine then a “?” and various parameters but including the search term itself. Again, depending on context, some character substitutions are made, such as “+” for space. The precise search term will depend on whether the search is on one or all words in the search term or an exact phrase.

Cookies

In addition one can see ‘cookies’. These are small pieces of information passed between the browser and the web site to maintain information about the session such as the contents of a 'supermarket trolley'. A feature of these is that they have names which will generally correspond to the web sites issuing them. Presence of a cookie indicates that the related site has been visited.

Deleted files and Free or Unallocated Space

Material deleted from a Windows PC by a user goes first into the ‘Recycle Bin’. This is essentially a folder (directory) with special properties. Deleted material in the Recycle Bin is intact and completely recoverable. The main distinction between the recycle bin and a normal directory is that it may be cleared from the recycle bin at the system’s discretion. A fixed-size area is reserved for the recycle bin and depending on use, material could remain there for a very long time.

When the system requires space in the recycle bin (e.g. for newly deleted material) it will clear out old files. At this point they become essentially lost to the system and will be overwritten when the area of disk surface they are on is reused. Some system processes bypass the recycle bin and return unwanted file space directly to the free space. All system information about material in the free space has been lost including the original file name and dates it was created, modified or last accessed.

A different case occurs when a whole folder is deleted. Then, though the files are returned to the free space as above there is enough information in the associated deleted folder to recover names and dates.

In all these cases the original content of the file remains intact.

A naïve user would not be able to recover information from the free space and would probably be unaware of its existence. There is actually a lot of software ‘out there’ that can recover deleted files and most is easy to use but the knowledge of its existence, understanding the need for it and knowing where to get it is beyond that of a naïve user.

As no information is attached to individual files in the free space very little can be concluded about them. Their origin cannot be inferred directly nor when they were downloaded or deleted - they aren't even files any more, just data. All that can be said is that this data once existed as files on the hard disk at some point and have since been deleted.

Pop-ups

Pop-ups in the context of criminal cases are unsolicited web pages thrown at the user as a result of browsing a web page that contains concealed code that generates them.

It is possible to embed controls in a web page that launch ("pop up") one or more fresh windows and *in extremis*, a storm of pop-ups can bombard the viewer at a rate that is difficult to control. Each of these will leave a permanent record in the web cache and can possibly add entries to 'favourites' lists and so pop-ups have become a common attempted defence in cases involving web abuse.

The mechanism is as follows:
One web site has arranged some affiliation with another. A typical example might be that a site specializing in one kind of pornography has links with another that specializes in something related but different. Unsolicited pop-ups are a nuisance so reputable sites generally avoid them but visiting any site to do with glamour or pornography can often result in a pop-up advertising further pornography or gambling.

There are limits on what can be ascribed to pop-ups. Generally they will be advertising banners as their function is to engage interest in the affiliated site. They will not be detailed, un-labelled images for the same reason and neither will they come from deep within a web site. They will not show the most extreme content. Also their scope will be limited to a handful of sites related to the site visited – if the cache shows systematic visiting of site after site this cannot really be attributed to pop-ups.

Peer-to-peer (P2P) systems – KaZaa, Gnutella, Limewire, Ares, torrents etc.

P2P systems are a class of programs commonly used for sharing material on the Internet and there must be millions of copies of such programs 'out there'.

The point of P2P systems is that each participating P2P agent (software) is indeed a 'peer' of every other and can both send and receive data and can participate in the discovery or indexing process.

During installation of the software an area called the shared folder (or shared area) will be set up and which the P2P agent will use to store material downloaded from the network. Once the installation is complete the user may enter search terms which are compared against what is known of the contents of shared folders of other compatible P2P users around the Internet. Matches are displayed in a table and the user may then click on an entry at which point the P2P agent on the user's PC contacts the agent on the remote PC and the file is fetched.

Searches are made on the file name, a point which will be returned to later. Files may contain anything but the majority of files shared are audio tracks, photographs and video clips.

An aspect of all file-sharing applications is that by default the shared folder is set up to be potentially visible to other users. 'Potentially' because it will only be made visible by the 3rd party typing in a search term that matches one of the files. It is not thrust at every other P2P user – someone has to make a matching request.

Files are only downloaded at the request of the user so files could only have reached the shared folder of the local PC at the explicit behest of someone using it. The corollary is that material in the shared folder could only have been sent out (uploaded) to the Internet at the specific request of a 3rd party elsewhere requesting it on the basis of some search term.

Sharing can be turned off at which point the P2P agent becomes a download-only program. This is deprecated by P2P communities and the user has to take specific action to do so. With some P2P system it is not even possible to turn uploads off.

The presence of a functioning copy of a peer-to-peer agent immediately provides the means to distribute (or perhaps more fairly stated, the means by which files on the user's PC may be exposed for upload by others).

The intent to distribute is difficult to gauge without other evidence. If uploading is turned *off* then there would appear to be an intent *not* to distribute but most P2P software does not allow this.

The file name is usually the only information available to anyone searching for or downloading it prior to viewing the image itself. Searches are made on the file name and for this reason P2P files names are usually highly descriptive. As well as private individuals sharing materials, P2P networks are used by commercial concerns who want to entice users to use their web sites. Such people will put just about anything in an image file name so that it will appear in a wide variety of searches. Other file names contain descriptions with are internally contradictory or contain as many related words as the author could fit in. The quality of the descriptions of material on other PCs is therefore highly variable and descriptions are often misleading. Lists of such names must therefore be regarded as circumstantial unless the actual image files can be located.

Another feature of P2P systems is that the quality of remote servers is highly variable – if one requests a file there is little guidance as to whether the machine it is being requested from is fast, small, busy or quiet or is on a slow or fast line. Also for the reason above among others, the file that is downloaded may not be what is wanted or may be of poor quality or the transfer may stall or go impossibly slowly.

For this reason it is customary to launch many requests at once and monitor their progress. Transfers that stall can be killed or suspended and resumed later, resulting in **incomplete** transfers that clutter up the shared area. Once the download has progressed a little it can often be **previewed** to see if it looks promising. If not the transfer can be killed.

md5 checksums

P2P systems *(q.v.)* need to be able to identify files uniquely regardless of different filenames etc. They do this by computing *md5 checksums* **or** *hash values* for files. The hash value is a large number produced by applying a mathematical formula to every piece of data in the file in turn and is so large and the formula so chosen that the probability of two different files generating the same hash value is microscopic. KaZaa for instance computes and retains the hash values of files in “My Shared Folder” so that if a remote KaZaa user wants to know whether or not the local KaZaa has a particular uniquely described file, all it has to do is send over the hash value. The local KaZaa compares it to its own list and can reply accordingly.

The police have a large database of hash values of known paedophilic images. Where they have found a match they can be confident that an exact replica of that file was present on the PC being examined even if the file itself is no longer present..

'Making'

The term 'making' appears in the Civic Government (Scotland) Act 1982 section 52 to do with child pornography and requires demonstration of some making process. From experience this is generally taken to relate to either:

1. Copying from one location to another
2. Searching out material and then saving it to permanent file space
3. Creating an image where none had existed before (e.g. via a camera).

The position seems to have tightened up in recent years and generally:

If material is found on removeable media (CDs, DVDs, floppy disks) with a reasonable expectation that it was copied there by the user (as opposed to having been acquired from a 3rd party) then making will be pursued under 1.

If there is a reasonable case that material was downloaded the making will be pursued under 2. , even if the images found were in the Internet cache. The argument here is that following browsing a copy existed where none existed before.

Situation 3. would appear to be clear-cut.

Charges of 'making' seem almost automatic except where images are found in a location with no clue as to how they got there (e.g. on a non-networked PC).

The following determination was recorded in August 2002 (http://www.sentencing-guidelines.gov.uk/docs/advice_child_porn.pdf):

SENTENCING PANEL'S ADVICE TO THE COURT OF APPEAL ON OFFENCES INVOLVING CHILD PORNOGRAPHY

23. We also proposed in the consultation paper that the downloading of indecent images onto a computer for personal use should be treated, *for sentencing purposes,* as equivalent to possession, despite the Court of Appeal's decision in *Bowden*[3] that someone who has downloaded such an image may properly be convicted of 'making' an indecent photograph under section 1(1)(a) of the 1978 Act. Our reason for this was that 'making' in the sense of making or taking an original indecent film or photograph of a child is clearly a more serious matter than downloading an image from the Internet, which is more akin to buying a pornographic magazine from a shop or mail order service. The majority of our respondents agreed, and this is the line we follow in our advice.

24. A more recent Court of Appeal decision[4] has further extended the interpretation of 'making', to include a simple request for the downloading of an indecent image so that it is displayed on screen. It is no longer necessary for the offender to take any further action

to ‘save’ the image, although the prosecution does have to prove that the accused knew what sort of image he was calling for. The effect of this judgment is that a conviction of ‘making’ can be based solely on the locating by a computer expert of an image in the Internet browser ‘cache’, provided there is additional evidence to show that the offender was seeking such material. The Panel suggests that the starting point for sentence should be lower in such a case than in one where the offender has actively saved the material.

PC file dates

Files on Windows carry three dates, the 'created' date, the 'last written' or 'last modified' date and the 'last accessed' date.

The created date is when that instance of the file was created.
The last written date is when the content of that file was last altered.
The last accessed date is when that instance of the file was last read or written

Thus:

If the create date precedes the ‘last written’ date the file was created then altered *in situ*.

If the ‘last written’ date precedes the create date the file was copied from elsewhere. Note the distinction between the creation of the *content* and copying to create the *instance* of the file seen.

If the created and last-written date/times are the same, the files were created *in situ* where seen or *moved* there and not altered subsequently.

Copying a file creates a new instance of the file with a new create date but with the original ‘last written’ date. Moving a file shifts it to a new location but leaves both dates intact. Any copy or move operation alters the last accessed date to when the operation occurred.

Files on CDs carry only one meaningful date - the 'last written' date since the create and last accessed dates will automatically be the date the CD was burned. The burn date is contained in the CD volume header and can be displayed using programs such as 'Nero'. For reasons of compatibility this single date is sometimes reported as the create date.

Chatrooms and social networking sites

The Internet supports a whole range of systems which allow users to communicate in "real time" i.e. conversationally. Communication can be one-to-one, one-to-many or many-to-many and conversation can be by text, sound, video or a combination of some or all of these.

In order to make conversation possible between the possibly millions of chat users, chat systems are divided up into 'chatrooms' identified by a topic or an on-line personality. A chatroom will support typically up to 30 or 40 people at once, more if they are operating one-to-many. Over this number chat becomes impossible as there are too many crossing conversations. Users rarely identify themselves by their real name and nicknames or graphic "avatars" are used – there is therefore a high degree of anonymity and chatroom users engage in personal conversations or explore forms of behaviour and relationships that they would never contemplate in the real world.

As well as conversing in the chatroom most systems allow private messaging at the same time. Two people can therefore join the conversation at large while chatting privately.

Some chat systems allow a "friends" list. Each user's chat login status is reported to all the others and they will be notified by a message when a friend logs in. This can be disabled by 'blocking' a friend – usually a sign of a fairly major falling-out.

When the user first logs in to a chat system he/she will be shown a list of active chatrooms, possibly with status information such as the number of users and maybe their nicknames. On logging on his screen will show a split window. One pane will show any message the user types as it is being composed. A second will show messages being sent in to the chatroom by other users in the order they are received by the chat server. A third may show a list of users. Users can usually be selected for a 1:1 message or alternatively if they are being obnoxious filtered out. Some systems allow a user to be invisible to some or all other users.

Video-based chatrooms tend to run with chat and video (sound has limits and is surprisingly heavy on resources). Video windows are small, maybe 5 * 4 centimetres and most run only a few frames/second. The purpose of the video is to give a sense of who one is talking to, not for verbal communication or much by way of expression.

Though there are perfectly serious applications for video chat such as distributed classrooms, it also lends itself to eroticism and many video chatrooms have the feel of fairly wild pyjama parties with people shedding some or all clothing and engaging in sex or masturbation, all on camera. Quite frequently one chatroom member will 'hold court' and many systems will have a chatroom with a name such as "Nikki's room" where "Nikki" will be the focus of attention by force of personality or other attributes.

Because many chatrooms are sexually charged and because of the anonymity they are a hotbed of internet relationships of all kinds. These range from plain friendships through flirting through exploration of alternate sexuality to significant relationships that can be constructive or destructive. Marriages have occurred or been destroyed as a result of chatroom relationships and anyone engaging in an internet relationship has to bear in mind that the person they are talking to may bear no resemblance in the real world to their appearance on chat. This is how predators operate and there are quite rightly concerns about the degree of access of minors to such chatrooms.

Particular care is required for fairly obvious reasons when a chat relationship jumps into the real world. A relationship will develop with progressive revelation about who the parties are and where they live etc. as trust builds up but predators are well aware of this and act accordingly.

For speed and to allow representation of emotions, chat conversations have their own language with some elements in common with phone text messaging

lol	Laugh Out Loud
b/f, g/f	Boyfriend, girlfriend
rotfl	Roll on the Floor, laughing
afk	Away from keyboard
brb	Be right back

and so on.

Chat may be embellished with graphics such as a laughing or crying face, a thumbs-up sign or many other ‘emoticons’ or ‘winks’. Nicknames can be as short as a single word or a full line of text representing the owner’s state of mind at the time. Nicknames can be changed at will during a conversation and often are. Chat is ephemeral and may not be stored on the PC more than momentarily unless explicitly saved. Precise behaviour depends on the chat system.

Social Networking sites – Bebo, MySpace, YouTube

These extend the idea of a chatroom and combine elements of chat, personal web space, on-line diaries or ‘blogs’, multi-user contributed web pages or wikis, email, streaming download of audio or video and so on. They are enormously popular amongst the 13-25 age group and Bebo for instance claims over 80,000,000 users.

Copine Scale

A scale defined by the Combating Paedophile Information Networks in Europe, at the University of Cork

Level	**Description**
1	Images depicting nudity or erotic posing, with no sexual activity
2	Sexual activity between children, or solo masturbation by a child
3	Non-penetrative sexual activity between adult(s) and child(ren)
4	Penetrative sexual activity between adult(s) and child(ren)
5	Sadism or bestiality

Applied separately to pubescent or pre-pubescent children

Cf: also 10-point Copine scale and BBFC guidelines

Copine 5	Copine 10	Description
1	1	Indicative
	2	Nudist (naked or semi-naked in legitimate settings/sources)
	3	Erotica (surreptitious photographs showing underwear/nakedness)
	4	Posing (deliberate posing suggesting sexual content)
	5	Erotic posing (deliberate sexual or provocative poses)
	6	Explicit erotic posing (emphasis on genital area)
2	7	Explicit sexual activity not involving an adult
3	8	Assault (sexual assault involving an adult)
4	9	Gross assault (penetrative assault involving an adult)
5	10	Sadistic/bestiality (sexual images involving pain or an animal)

Newsgroups

The term Internet tends to be used as a synonym for “World Wide Web”. Actually the Internet is a carrier for the Web and a great number of other services. The Internet is much older than the Web and its origins can be traced back to 1968 as opposed to 1990 for the Web. *Usenet newsgroups* were one of the earliest forms of multi-person communication.

A newsgroup is basically a location where anyone can post material for all to read. A posting of value will tend to generate responses and if these maintain a topic they are called ‘threads’. Newsgroups are organised so that the topics are kept within bounds tight enough for meaningful discussions to take place.

For this reason, newsgroups (of which there are over 100,000) are named hierarchically, starting at the top with “comp.” (computers), “news.” (news), “rec.” (recreation), “sci.” (science) and a few others, most notably “alt.” (alternative).

The naming scheme then proceeds with topics and subtopics separated by “.”.

Examples:

comp.infosystems.www.authoring.tools	Programs to help authoring Websites.
comp.infosystems.www.browsers.mac	Web browsers for the Macintosh platform

Usenet groups form the backbone of discussion-based information on the Internet and have done for many years. At one end of the scale they are a vital technical resource, a source of academic information or a vehicle for free speech. At the other extreme they form an outlet for a myriad of strange ideas, cults, fetishes and a source of on-line raw material of every conceivable form.

Usenet discussions are not restricted to text any more than is e-mail. Usenet groups created for posting of non-text information tend to have the word ‘binaries.’ in the name with some indication as to what the binary information represents (e.g. pictures). Like much else on the Internet there is no Usenet central authority so anything goes. Reputable sites restrict the groups they are prepared to support but whatever the information, someone somewhere will be prepared to host it.

Because of its historical origins, the Usenet mechanism is quite primitive by current standards though very efficient and effective and usually Usenet groups are accessed by mechanisms that add functionality and provide extra facilities. Usenet newsgroups can be read with dedicated software or through a Web-based service which allows them to be viewed with a web browser. Applications and services will have added features such as indexing or thumbnail-based previewing of pictures.

On first encountering Usenet, Users are presented with a list of groups and subscribe to the groups of interest. This is simply a process of informing the application or service about which groups are to be checked and does not imply that money is being paid. (Most Usenet postings are free of charge though the hosted service may be charged-for)

Usenet groups are dynamic as they are bulletin boards not web sites so material tends not to stay on news groups for long. For this reason the assumption is that material will be downloaded to a user-specified area for off-line perusal rather than being browsed *in situ* (though this can be done via hosts).

Because the newsgroups are already hierarchically named and because there will often be continuity between one posting and the next (e.g. discussion threads) it is usual to store/archive Usenet material sorted by news group.

Computer-Forensic Methodology

Computer-Forensic procedures in the Western world follow well-established principles intended to preserve all evidence originally present and at the same time guard against accidental or deliberate corruption of or tampering with the evidence.

In the UK these are enshrined in the ACPO guidelines for handling of computer evidence (http://www.acpo.police.uk/asp/policies/Data/gpg_computer_based_evidence_v3.pdf) and include the following:

> A computer under investigation is never switched on with the hard disk in place. Instead the hard disk or disks are removed from the PC and examined independently.
>
> A certified write-blocking device is employed at all times any PC hard disk or other recording media is under investigation. This prevents any possibility of contamination of the evidence.
>
> The first thing that is done once sufficient photographs and/or noting of serial numbers has taken place is that a forensic image is made of any recording media present, including hard disks. Again, this is performed using known and accredited tools (usually a program called EnCase).
>
> The forensic imaging process reads the disk at a very low level i.e. what is made is essentially a true replica of what was on the disk. This faithfully records files, deleted files and all residual data in all forms on the disk. A corollary is that an identical copy of the hard disk contents can be created if required.
>
> The imaging process also includes the generation of a hash code, a large number created by applying a mathematical formula to each piece of imaged data in turn. This number is appended to the image file - any attempt to alter the imaged data after this point can be checked by re-calculating the hashing algorithm and comparing it to the stored copy. Any alterations to the imaged data would produce a different hash code from the original.

It can hopefully be seen from this that it is very rare for the hard disk itself to be required in a defence examination unless there was suspicion that the evidence had been tampered with. In all other cases the forensic image is an entirely satisfactory copy of the hard disk contents. The forensic imaging process is slow and a complete independent examination would require something like an hour per 20Gb simply to image the disk (i.e. 3 hours for a 60Gb disk) just to get started.

The most common kinds of subsequent examination are then to locate files containing photographs (both intact and deleted) or to locate particular character strings (names, addresses, credit card nos. etc.) wherever they occur on the disk. Both tasks are highly automated but can again be slow (an hour+ per function).

Viruses and Trojan Horse programs

A very common defence is to cite Trojan Horse programs and viruses as possible causes of undesired activity. The most practical way of checking this is to reconstruct the hard disk contents in its original form then run a comprehensive collection of commercial anti-trojan etc. software on it. It should be noted however that Trojan Horse software, though often capable in principle of many forms of intrusion are generally used

a) to spy on data to recover passwords, credit card details etc.
   or
b) to lie in wait for instructions then wake up and run the PC as a ‘zombie’ process which is then used among thousands of others to (typically) send out SPAM or flood a victim machine with data to bring it down.

Thumbnails

Thumbnail images or thumbnails are small images displayed on a web page to aid navigation. They are typically an inch square or thereabouts when displayed and just enough detail can be made out to give an idea of what will be encountered when they are clicked-upon. An example might be an art gallery web site – 10 or 20 thumbnails would be displayed on one page and would give an immediate idea of whether the images they represent are landscapes, portraits, abstracts etc. They would not be detailed enough to determine much about the image and would not be something one would spend much time looking at.

Analysis of the browser *cache* following browsing of a web site containing thumbnails would show all the thumbnail images in some cases accompanied by the equivalent full-size image. The conclusion would be that where the thumbnails were not accompanied by the full-size image, either the person browsing the site was not interested in their content or the full-size images had faded away or been deleted from the cache. Where the cache index is still intact this can probably cast more light on exactly what happened.

Where numbers of images are quoted in forensic reports it is important to ascertain whether thumbnails are included in the total as there is an argument for excluding them on the grounds that they are part of the navigation process. The user may also not have much warning of their content before they are in view at which point they will have been copied to the web cache.

If the thumbnail is copied out of the cache then there is a counter-argument that it is no longer a thumbnail and should be regarded as a normal image, albeit a small one.