9. Forensic artifacts

This chapter will discuss analysis of smaller byte patterns found on a drive. First starting with file fragments, it will discuss how to piece them together to reconstruct files. Then it will discuss artifacts like email address, phone numbers, Web links, and similar data. These can often be found by scanning a drive without necessarily using its operating system or Master File Table, by just treating the drive as a sequence of bytes. (Boddington, 2016) provides a good introduction to how these artifacts can be used in criminal investigation.

9.1. File fragments

Files may be stored in pieces as part of the normal activity of an operating system. This fragmentation can be due to a file increasing unpredictably over time, as with log files that record important events for software when periodically more storage is needed beyond its allocation. It can also be due to a drive getting close to full and needing to fragment files to fit them into the remaining unallocated storage. Most often however, it occurs when files get marked as deleted and start getting overwritten with other files; usually, only parts of the deleted file get overwritten because it is rare that new data is exactly the size of the old data. 2.5% of the undeleted files in our corpus were fragmented, but rates for the author's drives were higher because of the big-data applications run on them. Table 1 shows some example counts of fragments in our corpus. Many of these were big; big files are more likely to be fragmented.

Table 1: Number of fragments in our corpus for some sample files.

Name of file	Average number of fragments per drive	Total size
W_M_Hair_024.mod	10	92329
2823.txt	3	8475
SCAN.CHM	5	327839
37c2890e6c4f[1].jpg	2	6537
DelDups.dll	4	176128
wmc_bw120.jpg	2.1	5339
35310.txt	2	6393
MSId0f7f.LOG	2	8586
259cf.msi	44	4075520
A0205783.sys	31	453632
padraig_10.jpg	3	266330
US_0_name_massey_01.wav	3	107892
cleardot[1].gif	37.8	329553
A0036260.ttf	2	29680
M_B_Special_Upper_03.mod	18	148427
level105.lev	4	15854
A0051995.dll	34	303104
NSRulerView.h	4	12820
nusrmgr.cpl.zottel	5	294912
MSIae29.LOG	2	8588
A0054747.exe	24	231288

Links to fragments of an undeleted file are stored with the master file table. A forensic image-analysis tool must work as the operating system does to assemble those fragments into a whole for a user. Fragments of deleted files are often not tracked. Nonetheless, the frequent duplication of files by an operating system and users means that copies of missing fragments can often be found on a drive. They can often be assembled by noticing overlaps or checking clues to consistency. For instance, words of fragments of a text file can be matched between fragments to find plausible continuations. With many formatted files, special characters at the front and back of the file indicate the first and last parts of the file; for instance, JPEG-format image files always start with hexadecimal "FFD8" and always end with "FFD9". JPEG also has a segment of EXIF ASCII data describing the source and conditions of creation, and also has a color table, and both can be identified in fragments.

Using clues like these, we can piece together file fragments in a process called "file carving" analogous to assembling a jigsaw puzzle. If the operating system has multiple copies of files, the missing parts of one file may be found in another copy elsewhere on a drive. And even if some parts of a deleted file cannot be found, enough may be found for investigative purposes, and may suffice in a criminal investigation.

Quick Question 9-1: Assemble the following fragments into as much of a coherent message as best you can, guessing the missing parts, and assuming this was a message sent by a spy: "oward Baghd", "rting 0700 from" "forward base Del" "sual precaut", "arge convoy".

File carving methods do not work as well on flash-memory storage since often whole blocks of storage must be erased before being overwritten. However, block erasing is slow, and modern methods apply garbage-collection techniques to postpone erasures until a good time occurs to do many of them at once.

9.2. Personal artifacts

Several kinds of personal data can be found on most drives, such as email addresses, messaging names, phone numbers, street addresses, and personal names, and these can be quite useful in investigations. We call these *personal artifacts*. These can often be found without opening files since they are often stored as ASCII or Unicode strings. Many tools like Autopsy will scan for these in images when requested. Search often uses "regular expressions" to define what it is looking for, strings with several kinds of special characters that permit matching to anything in a class of characters. For instance, email addresses have the symbol "@" preceded and followed by characters, numbers, and punctuation marks up to a maximum length. Delimiting of artifacts is usually represented by punctuation marks such as spaces or quotation marks, but other delimiters occur. For instance, for "<jsmith@hotmail2.com>", the "<" and ">" angular brackets delimit the address in an XML format.

9.2.1. Email addresses

Email and messaging addresses are a high-quality source of information about the contacts and interests of a user of a drive. Finding them is considerably easier than searching through files for keywords because of the rare character "@" they have. Furthermore, the international standard RFC3696 for email addresses specifies no more than 64 characters in front and no more than 255 characters after the "@", plus some restrictions on punctuation marks, which rules out many non-email uses of "@". Although addresses may be encoded in document formats like Microsoft Word, many appear as plain ASCII in Web downloads, log files, and software, and they are easy to collect there.

Our corpus had widely varying numbers of email addresses per drive. It had 292.3 million addresses in 2401 drives, 17.5 of which were unique. 61.3 million files occurred on these drives, an average of 4.8 unique addresses per drive. On the other hand, an old drive used in our office from 2007 to 2014 had 1.2 million email addresses, of which only about 50,000 were unique, as many occurred repeatedly like certificate-authority addresses.

A key problem with email addresses is that many found on a drive occur in software as contact addresses, and should be excluded from most investigations. For instance, a criminal investigation will not be interested in support@microsoft.com, a contact address for question answering. We can make a list of common uninteresting addresses, a "stoplist" analogous to those we use for searching in text files. Some examples from an email stoplist from NIST's scan of their software files are given in Table 2. Almost all were contacts for the software, though some of those were actually personal email addresses.

Table 2: Sample of a stoplist of email addresses (uninteresting addresses for investigation).

lpiiiytte@ytte.de
lpilon@your.domain.name
lpinto@ee.fit.edu
lpinto@theos.com.mx
lpitcher@sympatico.ca
lpitcher@yesic.com
lpitta@scuacc.scu.edu
lpj@ans.net
lpk@cs.brown.edu
lpkruger@flagstaff.princeton.edu
lpkruger@phoenix.princeton.edu
lplanas@ya.com
lple@us.ibm.com
lplqmx@vol.vnn.vn
lply@jw.ki
lpm102@psuvm.psu.edu
lpm@leox.org
lpm@mirth.demon.co.uk
lpmaniccia@aol.com
lpmeissner@msn.com
lpn-l@brownvm.brown.edu
lpnw@ximian.com
lpodlipec@wellfleet.com
lpoetter@src.gnome.org

It is impossible to give guaranteed rules for which addresses are interesting and which are not in an investigation. Instead, we can combine clues in a potential email address to rate the likelihood of it being interesting, much as we combined clues to rate the likelihood of malware occurring in a particular place in a file system in section **Error! Reference source not found.**. Example clues f or addresses are whether it was found in software, the number of drives on which it occurred, whether preceding characters suggest software, the domain type given after the "@" (especially if it is a server), and recognizable user-name words.

Clues can be tested systematically. Figure 1 plots two examples, the probability of a forensically interesting email address as a function of the length of the user name (blue curve) and domain name (green curve). This data came from a training set of 7638 randomly selected addresses from our corpus, where we manually tagged the items based on a little research. Ten characters appears to be the most popular length of a user name as a compromise between being easy to remember and providing sufficient variety. However, some long user names were automatic forwarding addresses using near to the maximum permitted length. Users do not usually have a choice with domain names, but most of them were under 20 characters.

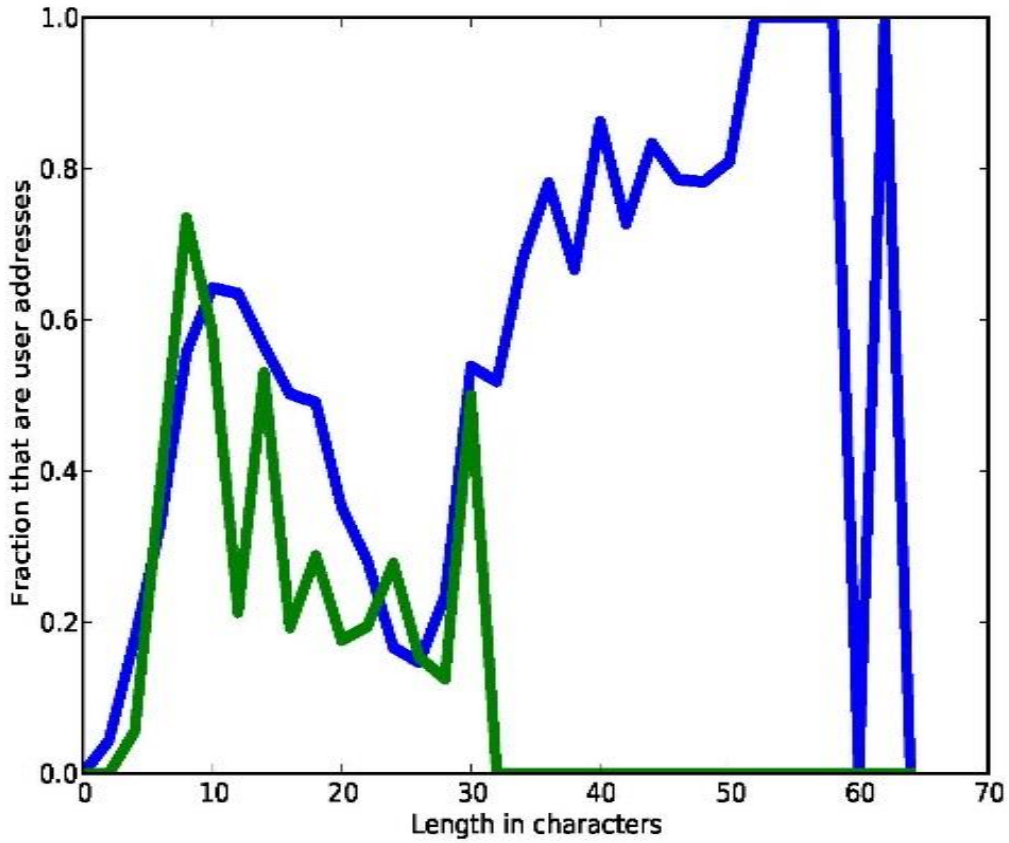

Figure 1: Probability of a forensically interesting email address as a function of the number of characters before the "@" (blue curve) and number of characters after the "@" (green curve).

A good negative clue to an uninteresting email address is its number of successive occurrences in the sequence of all addresses occurring on a drive (Figure 2). Large values of this number probably come from log files, and are almost certain indicators of uninteresting automatic activity.

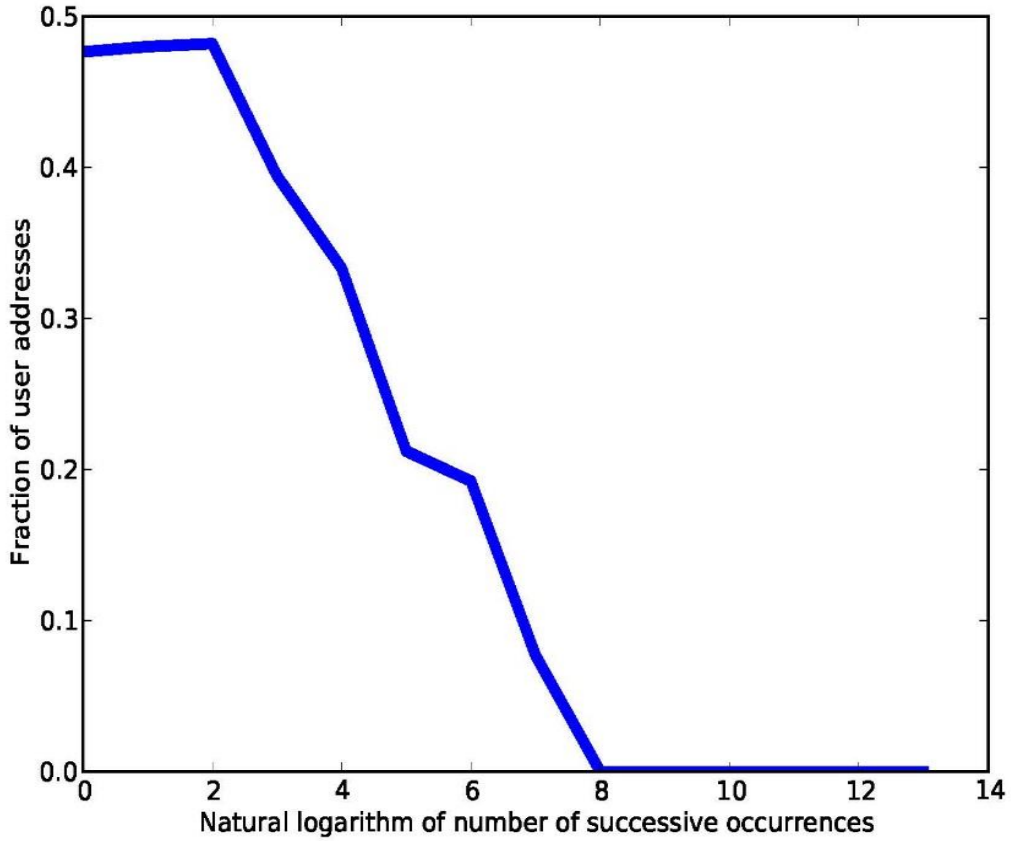

Figure 2: Probability of a forensically interesting email address as a function of the number of successive occurrences in all the addresses on a drive.

Table 3 shows the odds for some other interesting-email clues based on our random training set. As before, odds are defined as the probability divided by one minus the probability. We can use these clues in a Naïve Bayes calculation like that in chapter 5 to estimate the overall likelihood of an address not previously seen as being interesting. Note these calculations require some word lists for software terms, mail servers, and areas of the world. The formula would be as follows, where "o" means odds, "I" means "interesting", "o(I|X)" means the odds of I given X for some X, and odds is the probability divided by one minus the probability:

$$o\big(I\big|(C_1 \& C_2 \& \ldots \& C_N)\big) = o(I|C_1)o(I|C_2)\ldots o(I|C_N) \,/\, (o(I))^{N-1}$$

Table 3: General clues to interesting email addresses.

Mean odds	Description
0.116	Address in stoplist
1.035	Address not in stoplist
0.078	Software-suggesting preceding characters
1.038	No software-suggesting preceding characters
1.230	Occurs only on one drive in the corpus
0.556	Occurs on two drives in the corpus
0.281	Occurs on 3-10 drives in the corpus
0.234	Occurs on >10 drives in the corpus
0.583	Domain length < 8
1.364	Domain length 8-15
0.274	Domain length > 15
3.284	Server name in domain
1.350	.edu domain
0.076	.org domain
0.519	.net domain
0.106	Other domain
0.063	A username word matches a domain word
1.054	No username word matches a domain word
1.225	U.S. domain
0.327	Developed world non-U.S. domain
2.582	Developing-world domain
0.465	Other domain
0.245	Username < 8 characters
1.355	Username 8-15 characters
0.802	Username 16-29 characters
3.583	Username > 29 characters
0.886	First username character is digit
0.887	First username character is not a digit
2.149	Last username character is digit
0.693	Last username character is not a digit

Our training set had 7638 entries of which 3190 were identified by us as valid interesting email addresses. Using a Poisson model, the standard deviation of the count of a totally random feature of the addresses should be 56.5, which means that three standard deviations above and below the mean (a standard statistical criterion for significance) should be probabilities of 0.440 and 0.395, which correspond to odds of 0.785 and 0.654. All the odds in the table are outside this range except for the last, which therefore represents a clue that is not significant and should not be used.

Using the Naïve Bayes formula and converting resulting odds to probabilities, some estimated probabilities that our method calculates for sample email addresses are shown in Table 4. In

general, most ratings are near 1 or near 0 with only a few intermediate ratings for addresses with conflicting clues (Figure 3).

Table 4: Example email addresses in our corpus and their calculated probabilities of being forensically interesting from a Bayesian model.

Email address	Calculated probability of being interesting	Major negative factors
support@b3d.com	0.0001	Generic username
premium-server@thawte.com	0.0001	Generic username
anyuser@zap.co	0.0001	Generic username
orders@amazon.com	0.0002	Generic username
tnelson@guildmortgage.com	0.0033	Business domain
naftali@umi.co.il	0.0035	Business domain
3a84eb988e5e3a4cb64d6936a93a79e11d9c@server2.ms	0.0170	Artificial username
aramirez@diego.iner.gob.mx	0.0174	Government domain
last@toc.v7686333.sa	0.0300	Odd domain
95hongchai@mfcfund.com	0.0648	Business domain
bee_imm.yap@cpf.gov.sg	0.0652	Government domain
schulleit.sertuerner.real@freenet.de	0.5074	Odd username
0gem00kh4xtdc5@smail.emirates.net.ae	0.5149	Artificial username
david_o2@012.net.il	0.5150	Odd domain
ao_luck@hotmail.com	0.5156	Odd username
mittalgaurav.dtii@gmail.com	0.9004	
u003cmereena3paul@gmail.com	0.9004	
rahimtulla125@jeevansathi.com	0.9807	
vaishalibag19@yahoo.co.in	0.9907	

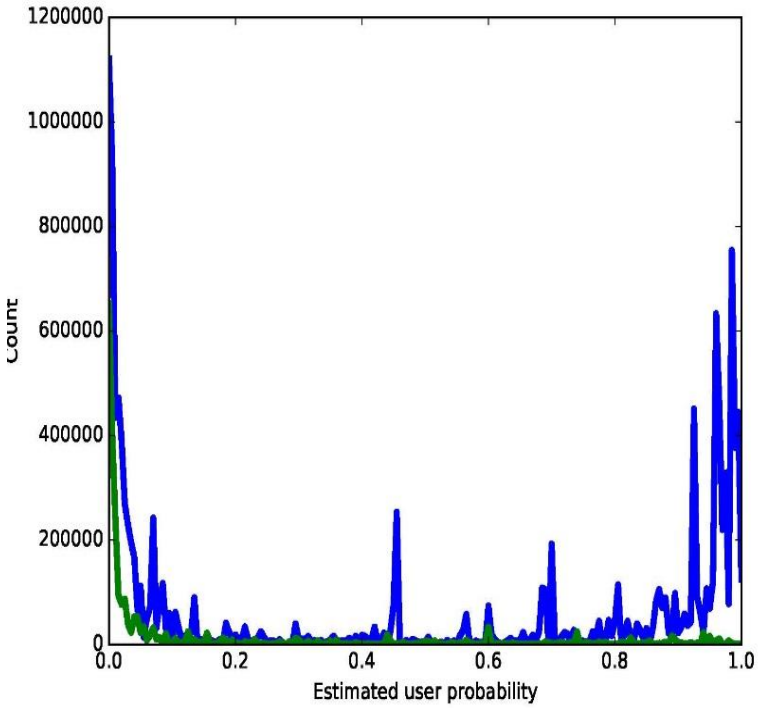

Figure 3: Distribution of interestingness ratings of email addresses for forensic investigations by our Bayesian method. The blue curve is for our corpus, and the green curve is for our stop list from NIST.

Users today tend to have multiple aliases (user names) for different systems due to the many large user communities for which users need unique names. Aliases are hard to find with forensics. They are unlikely to be nearby on the same drive because many of these addresses are used with separate software and their records will not be adjacent. Aliases tend to co-occur only on drives used by the same person, and these are rare, so co-occurrence is not a useful clue to them. That leaves only the clue of common words (like "john.h.smith" and "smithjh") to indicate possible aliases, plus a few rules of thumb such as initials for full words. A smart criminal will know how to avoid providing such clues with reusing names, and in fact, aliases dissimilar to the real name may be a clue to criminality, as in the quote at the start of this book. However, network forensics based on Web addresses (URLs) or Internet addresses (IP addresses) can track two aliases to the same source.

A complication of email addresses is that names of cookies (persistent data about Web use) can look like addresses, since they are usually small files whose name consists of a user name and a site with which they are associated. We discuss cookies more in section 9.2.3.

9.2.2. Phone numbers

Another common kind of data to extract from drives is phone numbers, and they can also provide useful information about personal contacts. Their formats are not specified by an international standard, and they are more varied than email addresses. Common formats (where "#" represents a digit) are shown in Table 5.

Table 5: Common formats for phone numbers.

Format ("# "means digit)	Interpretation
(###**)** ###-####	American general number
###.###.####	American general number
###-###-####	American general number
###-####	American local number
####	Phone extension within an organization
1-###-###-####	American general number for international use
########	International number for 2-digit country code, with optional spaces or hyphens between digits
#########	International number for 2-digit country cod, with optional spaces or hyphens between digits
##########	International number for 2-digit country cod, with optional spaces of hyphens between digits
#########	International number for 3-digit country code, with optional spaces or hyphens between digits
##########	International number for 3-digit country code, with optional spaces or hyphens between digits
###########	International number for 3-digit country code, with optional spaces or hyphens between digits

False positives (misidentified strings) occur more often with phone numbers than with email addresses because digit strings are used in many other ways in the digital world. For instance, IPv4 protocol Internet addresses consist of four numbers, each in the range 0-255, and IPv6 addresses consist of six numbers. These are usually separated by periods (".") rather than spaces or hyphens and so can be distinguished from phone numbers, but some people are now using periods for phone numbers too. However, the limits on the sizes of these numbers eliminate ambiguity in most instances.

9.2.3. Network artifacts

Another common type of data seen in many files are text Web links ("URLs") such as http://faculty.nps.edu/ncrowe/index.html. They are easy to recognize if they use an "http://" or "www." at the front, but these characters are optional. Microsoft Windows uses backward slashes ("\") to separate directories in the link path, but other systems use the forward slash ("/")'; forensic tools like Autopsy can usually find both kinds of links. Particularly important links are in the "recently visited" and "favorites" lists for a Web browser, so those are good places to check first in assessing the usefulness of a drive. If the end of the link does not give a file name with a period followed by extension like HTM or HTML, the default name is index.html.

Many investigators tend to overvalue these links found by scanning the entire drive. Bear in mind that links often occur in Web pages, and when a user visits a page, all its links will be cached on the user's computer or device, regardless of whether the user followed the link or not. This means Web caches often contain large numbers of advertising links that can be ignored. Furthermore, much software itself contains a URL link back to the vendor of the software, and this is not interesting in most investigations. This is why it helps to create a stoplist of commonly seen URL links.

The remaining potentially interesting links can often suggest user interests and associations. However, they interesting ones are not generally the ones most often found on a drive; Table 6 shows the most common links from a drive by the author used from 2005-2015.

Table 6: The most common URL links on an old drive of the author.

Count	URL link
93085	http://www.w3.org/2000/09/xmldsig#
85496	http://www.nps.navy.mil/Courses/msa847/ sasdoc/sashtml/common/images/cont1.gif
84752	http://www.nps.navy.mil/Courses/msa847/ sasdoc/sashtml/common/images/next1.gif
81956	http://www.nps.navy.mil/Courses/msa847/ sasdoc/sashtml/common/images/prev1.gif
57955	http://schemas.openxmlformats.org/drawingml/ 2006/main
42709	http://components.groove.net/Groove/ Components/Root.osd?Package=net.groove.Groove. ToolComponents.GrooveCommonComponents_DLL& Version=0
32335	http://schemas.openxmlformats.org/ officeDocument/2006/relationships
31102	http://www.mulax.info/Games-cheats.htm
30612	http://www.w3.org/2000/09/xmldsig#sha256
26677	http://oval.mitre.org/XMLSchema/oval-definitions-5#windows
25670	http://ns.adobe.com/xap/1.0/
25267	http://crl.microsoft.com/pki/crl/products/ microsoftrootcert.crl0T
25061	http://www.microsoft.com/pki/certs/ MicrosoftRootCert.crt0
23822	http://astro.berkeley.edu/~central/archive/us-cert
22840	http://g-images.amazon.com/images/G/01/ marketing/generic-promotion/2003/90/10-offer-90.gif

These are not too interesting since most seem to be automated records, except for the "game-cheats" reference for when the author was researching deception in games. However, if we look at a sample of the less-common links we see more interesting things (Table 7).

Table 7: A sample of less-common URL links on an old drive of the author.

Count	URL link
22	http://www.w3.org/XSL/Transform/1.0
22	http://www.wa.gov.au/perthobs/
22	http://www.wa.gov.au/perthobs/pics/big-logo.gif
22	http://www.wainwright.army.mil/4-123/bco/maintenance/dart/dart.html
22	http://www.wainwright.army.mil/4-123/bco/standards/areas/drive_train.html
22	http://www.wainwright.army.mil/mp/northern.htm
22	http://www.washingtonpost.com/wp-dyn/articles/A25105-2003Oct14.html
22	http://www.washingtonpost.com/wp-dyn/articles/A28446-2002Jan10.html
22	http://www.washingtonpost.com/wp-dyn/articles/A7786-2002Feb26.html
22	http://www.washingtonpost.com/wp-dyn/opinion/
22	http://www.washingtonpost.com/wp-srv/technology/interactives/ identitytheft/idtheft.html
22	http://www.washingtontimes.com/world/20020228-67420280.htm
22	http://www.washtimes.com/national/20020123-98656374.htm
22	http://www.washtimes.com/world/20040422-114403-9180r.htm
22	http://www.weather.gov

These indicate interest in news and some relationship with military organizations, which could be useful information in an investigation. Note in general that the sites listed may just be servers or proxies for groups of systems, or they may be deliberate front systems for obfuscation like the Tor "onion routing" software used for the "dark Web". So we may need to navigate to the real sites a user is using behind the ones listed.

You can also look for IP addresses and MAC addresses in scanning a drive. IP addresses in IP version 4 are four 8-bit numbers separated by periods, e.g., 133.83.20.119; traditionally the numbers are written as decimal numbers in the range 0-127. IP addresses in version 6 are eight 8-bit byte expressed as hexadecimal (base-16) strings separated by colons, e.g., 83f2:a77b:0038:9382:fffd:9e72:02c2:9472. MAC addresses are six 8-bite bytes, usually rendered in hexadecimal and separated by hyphens, and they refer to the machines in the local network. Mobile devices have IMEI and IMSI numbers (usually 15 decimal digits long) that identify them to mobile service providers, and which can help forensics. All these formats can be found by searches with regular expressions of a drive image. Generally they are not as useful as URLs since they give few clues by themselves about the purpose of the machines they reference.

Even more helpful in establishing the interests of a user are the queries they make to Web browsers. These are stored in different places depending on the browser, but most general-purpose forensic tools know where to look for them. The list below shows example Web searches found nine or more times on drives in our corpus. They do indicate particular interests of users.

- msbte
- yahoo.com
- tor
- rediffmail+login
- flash+player+download
- gmail.
- online
- shakira
- rani+mukherjee
- cache:yzZ7MosNOvsJ:video.google.com
- fo
- ss501
- world%20
- www.youtube.com
- live+cricket+score
- ork
- youtube.com
- madonna
- jobs+in+pune
- download
- office
- 4shared
- babes
- jessica+simpson
- java
- pune%20uni
- cache:tjm6a4o0xewJ:www.dhingana.com

- Golf+Vacation

A list of recently visited pages can also be useful in an investigation. Each browser has a specific directory in which they put copies (caches) of visited pages. These files are deleted only occasionally. Users may also explicitly download Web pages in which they are interested, and store them wherever they like, and these are treated like other user files and not deleted without explicit user direction.

Cookies can also help in investigations by identifying predominantly the commercial Web sites that a user has visited, and some data about their interactions with the sites. Usually cookies are stored as small files whose name is the username, the "@" sign, and the name of the site visited. Cookie files can include user names, addresses, customer numbers, financial details, and recently performed actions. Some cookies accumulate data about a user over several interactions. Some cookie data is encrypted, like bank-card numbers and phone numbers, but other information may be unencrypted and readable by investigators. Cookies are quite varied, and they may be worth checking in an investigation involving financial issues or unauthorized Web use.

Cookies are stored in a place specific to the Web browser. For instance, Google Chrome stores them in an SQLite database on Windows systems under AppData\Local\Google\Chrome\User Data\Default; Mozilla Firefox stores them under AppData\Local\Mozilla\Firefox\Profiles, Apple Safari stores them as plist files under Libraries/Cookies, and Internet Explorer and Edge store them for Windows 10 under AppData\Local\Microsoft\Windows\INetCookies.

Cookies can be for a session (like "shopping carts" that hold a list of items purchased in the session) or "persistent" (lasting after the user logs off, like those holding customer ID numbers, shipping addresses, and phone numbers to save time when the user next logs in). Persistent cookies can contain information for frequently used tools like Google. Such big persistent cookies can provide useful forensic data about where a user has been on the Internet. Cookies can also be "first-party" (created by a site for use on the site) or "third-party" (created by a site for another site, as with click trackers for advertisers). Since cookies are simple and most data is unencrypted, malicious actors can create malicious cookies to do bad things like sending malicious code to a site that sends purchases to the malicious actor instead of the user. Third-party cookies in particular are easily exploited. Thus cybersecurity scanning should check for suspicious cookies as well for malware.

Network use also leaves many cache files for temporary storage of data that is coming from or going to the network. They may get overwritten quickly depending on the software being used, but otherwise may persist for a while if their storage is not needed. An investigator may want to check them out. Video streams for example can take considerable time to erase, so you may be able to find their pieces on a drive.

9.2.4. Personal names

Another valuable kind of artifact found on drives are personal names since they can establish the people using a drive and their personal connections. Although email addresses may include personal names, and phone numbers can be looked up in directories, the occurrence of a personal name represents a less formal connection between people.

A limited set of words are used as personal names, so it helps to have a list of them. Table 8 shows some example names in our dictionary of 300,000 personal names from our online

appendix. These are primarily "given" names (the first names in English) but include some family names too.

Table 8: Example personal names from our collection.

joanann	joananna	joananne	joanathan
joandaly	joandra	joandri	joandry
joandy	joane	joaneil	joanel
joaneliz	joanell	joanelle	joanes
joanetta	joanette	joangel	joangela
joangie	joani	joanie	joanis
joanita	joanka	joanmarie	joann

Interestingly, a good name list has limitations. Several problems occur:

1. Many names found in drives are not in any listing of standard names, such as abbreviations and compressions of better-known names. Login names often have this characteristic because they must be unique for a system.
2. Many words in a language can function both as personal names and regular words. An example is "mark field" which is often found in JavaScript in our corpus as a command to mark a field, yet "Mark" and "Field" are common names in English.
3. Many names apply to more than one person in the world.
4. Names can be one, two, three, or more words long.
5. Capitalization in English conventionally indicates personal names, but digital data like file names are often in lower case.
6. Delimiters are not consistently used around names, so extracting the name itself can be difficult, particularly if it is only a single word.

Item 5 means that matching drive data to word lists should ignore case, but that does eliminate a valuable clue. Item 6 means that we should look for many kinds of matched delimiters, though quotation marks and spaces are the most common.

Item 4 requires some regular expressions. Table 9 shows regular expressions for some formats used in the U.S. These patterns are good clues to personal names because the likelihood increases with the more words involved.

Table 9: Some regular expressions for recognizing mulitword personal names.

given_name
family_name
given_name space family_name
family_name, space given_name given_name space middle_name_initial. space family_name
family_name, space given_name space middle_name_initial. given_name space middle_name space family_name family_name, space given_name, space middle_name
given_name_initial. space middle_name_initial. space family_name given_name_initial. middle_name_initial. family_name_initial.
given_name_initial. space family_name
given_name_initial. space family_name_initial.
given_name_initial family_name_initial

Quick Question 9-2: What minimum number of different people could be indicated by these names?

- Pablo Pedroza-Molina
- ppedroza
- pcp
- pedrona.pablo
- pabpedroza
- pedrozza_pablo
- pcmolina
- angrypablo

Item 1 on our list of personal-name issues can be partially addressed if we keep lists of dictionary words in many languages that can be excluded, such as words that function only as verbs and conjunctions in a language. Many email aliases are formed by combining two or more parts of names, so unknown words can be split systematically to see if we recognize the parts. However, these methods do not often help.

Identifying personal names is thus better handled with probabilistic methods to rate names on drives like with email addresses in section 9.2.1 rather than absolute rules. Table 10 shows the computed odds for several kinds of clues on a 5220-item training set we created. The probability of a personal name in this training set was $\frac{0.238}{1+0.238} = 0.192$, so we had 1004 positive examples and 4216 negative examples. Modeling this as a Poisson distribution, the standard deviation of the count would be 31.7, so the classic significance level of three standard deviations from the mean would correspond to odds less than 0.211 and greater than 0.266. This justifies us in discarding the clues not statistically significant: "> 13 characters", "single word", "multiple words", "only a personal name", 1-9 occurrences in corpus file names", "one occurrence per drive", and "1-5 occurrences per drive". Eliminating clues can often improve accuracy of a model such as this because the unnecessary extra clues can dilute the effect of the significant clues.

The owner or principal user of a drive is not necessarily the most frequent name on a drive, as often the most common names are those in frequent advertising. However, the owner name should occur frequently. We have found that setting a minimum threshold on the rating of a personal name can usually identify a small subset of names that are likely to be the owner or principal user. Clues can also come from the names in the Users directory in Windows or its equivalent on other systems, or the names associated with backup directories, but these can be aliases or nicknames.

9.2.5. Other artifacts

Several other kinds of artifacts can help in analyzing drives. The Bulk Extractor open-source tool (https://github.com/simsong/bulk_extractor) can find most of these if your main forensic tool cannot.

- Street addresses: These may be found with phone numbers and personal names in address books, but they take many forms and necessarily have a high false-positive rate. Nonetheless, we developed a program to find them with around a 40% success rate.
- Geographical location information: Often GPS (Global Positioning System) data is stored on phones, and it can localize the phone at a particular time. GPS data may also be found on wireless servers. KML is another geographical data standard often seen. However, a user may be at a different location than their phone, data may have been deliberately

changed to create an alibi, and the data may be from an imaginary world as in games. Figure 4 shows some example data found on the author's old drive; (36.5956, -121.874) are the latitude and longitude of his school.

- Bank-card numbers: These are rarely found on drives anymore since they are personal information that is a prime target for identity theft.
- Encryption keys: Those in AES format can often be recognized because they have a very limited set of lengths (usually 16, 24, or 32 bytes) and are delimited from their surroundings. Encryption keys could allow you decipher encrypted data, so you will not find them often.

Table 10: Odds of various clues to personal names.

Clue	Odds on training set
Length < 6 characters	0.167
6 to 13 characters	0.280
> 13 characters	0.226
All lower case	0.315
All upper case	0.138
Capitalized only	0.171
Mixed case	0.139
Delimited both sides	0.362
Delimited on one side	0.333
No delimiters	0.147
Followed by a digit	1.280
No following digit	0.210
Single word	0.234
Multiple words	0.244
Ambiguous word	0.058
Not ambiguous word	0.286
Only a personal name	0.238
Known name but not in corpus file names	0.533
1-9 occurrences in corpus file names	0.229
10-999 occurrences	0.174
> 999 occurrences	0.132
Normalized number of drives < 20	0.298
Normalized number of drives 20-200	0.450
Normalized number of drives > 200	0.114
One occurrence per drive	0.244
1-5 occurrences per drive	0.242
6-30 occurrences per drive	0.155
> 30 occurrences per drive	0.298
Organizational domain name nearby	0.017

No organizational domain name nearby	0.751
Before any clues	0.238

```
10852118528:    -07-04T21:02:59,51.4945,-0.148167,,,
54829170688:    T4/1 55/1 236/100,36.5977,-121.862,38.2307,,
68837449740:    2011-04-15T19/1 4/1 53/1,36.5956,-121.874,10,,
69763829760:    2014-03-11T22/1 5/1 5/1,26.3436,-81.7991,9,,
71629471756:    2011-04-15T19/1 4/1 53/1,36.5956,-121.874,10,,
102869901324:   2011-04-15T19/1 4/1 53/1,36.5956,-121.874,10,,
114587275264:   T4/1 55/1 236/100,36.5977,-121.862,38.2307,,
```

Figure 4: Example GPS data from the author's old drive.

- EXIF photograph metadata, described in chapter 8.
- Email and messaging headers. Although email is mostly stored on mail servers, you may be able to find downloaded mail on drives. Usually they follow the RFC 4322 standard which aids in searching for them. Full header information can show sites through which the mail was forwarded, which can be helpful against attempts to conceal the origins of email.
- Compressed files, e.g., zip and rar files: These often indicate downloads of related files.
- Files in data-interchange formats, e.g., XML and JSON. These often were transferred over the Internet.
- Headers of executable files, which are often follow standard formats such as PE (Windows executables), Mac-O (Macintosh executables) and ELF (Linux executables). This can contain useful information about the type of environment it requires and the basic parameters it uses.

9.3. Exercises

9-1. (*) The graph of the likelihood of an email address representing personal email as a function of the number of characters in its user-name portion showed that:

A. There was a maximum for around 11 characters and the likelihood became progressively smaller for larger numbers of characters.
B. There was a local maximum for around 11 characters and a local minimum for around 27 characters.
C. The likelihood consistently increased with the number of characters.
D. There were two big peaks for around 11 and 27 characters.
E. There was a peak at around 10 characters, then the curve went to zero at around 32 characters or more.

9-2. (*) What kind of information LEAST helps in deciding whether a word in a document is being used as a personal name when it is not in our list of known personal names?

A. The length of the word
B. The number and type of punctuation marks used near it
C. The number of occurrences of the word on the drive
D. The type of document in which the word was found
E. A dictionary of common words in many languages

9-3. (*) Suppose fragments of a Web page are left on a drive after it was marked for deletion and parts were reused for other files, but assume there were multiple copies of the page and we should be able to find fragments covering all the contents of the page.

(a) How should we determine the starting and ending fragments? Do some online research on the HTML standard.

(b) How should we rate the strength of a match between two fragments? Often a code repeats on Web pages, so matches may not be unique.

9-4. Suppose we find file deleted fragments of text (unencoded) log files in unallocated space on a drive. Log files record events that happened, when, and a few details. Suppose the only logs we expect to find on this drive are the system log of logins, the security log recording events such as updates, downloads, changing of security settings, and plugging in devices, and the network log recording basic information about packets sent and received over the network connection. What are the most important clues we can use to connect fragments together? What process should we follow to connect them? What must we calculate to predict when fragments are missing?

9-5. Consider the following fragments of an email conversation. Identify the most reasonable order to assemble them into a file message, and explain why it makes sense.

Fragment #1:

domex/papers/2010/proposal_fy12
On Dec 13, 2010, at 3:09 PM, Rowe, Neil (CIV) wrote:
Where do I find it?

-----Original Message-----
From: deep-research-bounces@lists.nitroba.org
[mailto:deep-research-bounces@lists.nitroba.org] On Behalf Of Simson

Fragment #2:

or, you can check out just individual files, e.g.:
svn co https://domex.nps.edu/domex/svn/papers/2010/proposal_fy12/proposal.tex

Fragment #3:

Garfinkel (CIV)
Sent: Thursday, December 09, 2010 6:59 PM
To: deep-research@nitroba.org
Subject: [Deep-research] proposal in SVN
I have committed the current version of the proposal to SVN. Please
email me if you want to see the spreadsheet.

Fragment #4:
From: Manley, James (Jim) (CIV)
Sent: Monday, December 13, 2010 6:11 PM
To: Rowe, Neil (CIV)
Subject: Re: [Deep-research] proposal in SVN

Hi Neil,
You have to check out files via svn co (or svn checkout), even if you're not going to modify and check them back in. Likewise, you need to use svn ls (or svn list) to navigate the repository, which is actually stored in a database and is not directly navigable via normal OS file commands. So, to see what's in the directory domex/papers/2010/proposal_fy12 (actually, domex/svn/papers ...):
svn ls https://domex.nps.edu/domex/svn/papers/2010/proposal_fy12

Fragment #5:
directory at any of the top levels. How do I get to it?

-----Original Message-----
From: Simson Garfinkel
Sent: Monday, December 13, 2010 3:38 PM
To: Rowe, Neil (CIV)
Subject: Re: [Deep-research] proposal in SVN

Fragment #6:
In any case, the repository directory structure (domex/svn/papers/2010/proposal_fy12, in this case) will be replicated in your local filespace, even if you're just checking out one file (and the directory structure will be populated with other files/directories as they may be checked out , in turn).
Hope this helps,
Jim

On Dec 13, 2010, at 5:10 PM, Rowe, Neil (CIV) wrote:

Fragment #7:
Simson

_____
Deep-research mailing list
Deep-research@lists.nitroba.org
http://lists.nitroba.org/listinfo.cgi/deep-research-nitroba.org

Fragment #8:

I would very much appreciate if everyone would review the proposal for typos or things that are not clear. This is the blueprint for research over the next 18 months. I'd like to have it be accurate.

Fragment #9:

Can you help get this file?
1. The svn "copy" command doesn't work -- it says that my destination address isn't a working copy, even when I give it the name of my local svn directory.
2. When I log into domex.nps.edu directly, I can't find any "domex"

Fragment #10:

and to check out (copy to your local filespace) the directory:

svn co https://domex.nps.edu/domex/svn/papers/2010/proposal_fy12

9-6. Examine the file sample_email_addresses below. For each item, discuss how likely it is a personal email address useful in investigations of people, and why; summarize your answers as "high", "medium", or "low". Consider the user name, domain name, and lengths of the parts as clues. It will help to look up the domains to see if you can determine which are mail servers.

prime_solutions@rediffmail.com
cbelbeze@capgemini.com
lisa.shaw@unisa.edu.au
srxirus@yahoo.com
df8eg9ln@uwiwykoqcgj7ei.jo
vyasfemale020-9890449197989044919 7sonal8vyas@gmail.com
balwan_singh11@yahoo.com
tutsaqq_@hotmail.com
jone189@gmail.com
nors@savion.cc.huji.ac.il
my.email.addr@comsoltech.com
benjamin.van.eeden@sap.com
serena_tan@rcb.gov.sg
quake@geophys.washington.edu
black_kingsley@yahoo.com

9-7. (*) Use the table of odds for clues to user email addresses in Table 3, and the formula on the previous page there, to estimate the odds that the email address "smithl482@gmail.com" is for a user. Assume it is not in the stoplist, is not preceded by software-suggesting characters, occurs on two drives, uses a server name in its domain name, and is a U.S. domain. Also use any other clues you can from the address except for the "username weight" clues. Assume the prior odds (odds in general) of a user address is 1. Then convert the total odds to a probability by the formula probability = odds/(1+odds). Show all your calculations.

9-8.(a) What are the main reasons that two personal names found near one another on a drive could be of unrelated people?

(b) What are the main reasons that two personal names that co-occur frequently on a set of drives could be of unrelated people?

9-9. One tool we use gets many false matches when scanning drives for bank-card numbers since it is looking for 16-digit numbers and they can serve many purposes. Give two contextual clues that would confirm that a number is a bank-card or credit-card number. Give two contextual clues that would confirm that it is not.

9-10. Suppose we want to anonymize all the email addresses, phone number, and personal names on a drive to publish an image in a public repository. But we do want the same item always replaced by the same anonymized string.

(a) What is necessary so that the length of each file on the drive remains the same after anonymization?

(b) What data structures would be best to track the anonymization?

(c) Under what circumstances must location information also be anonymized?

9-11.(a) When file carving with document text files, what formatting clues can you use to determine that one fragment is contiguous with another fragment?

(b) How could natural-language processing help you determine that one fragment is contiguous with another fragment?

9-12. IP addresses under IPv4 are the four-number strings delimited by periods that denote Internet sites. Suppose you collect all such addresses on a drive by a scan of its image. Give three clues that would help significantly in deciding whether an address would help in a malware investigation.

9-13. Suppose a file has been deleted, and part of its allocation in secondary storage has been written over with another file. Why could it still be possible to reconstruct the complete file using file carving, and how?

9-14. In the downloadable files are fragments of a text file with names starting with "splittext". You are to assemble the fragments into a reasonable text file. Describe the reasoning you use to determine which fragments are contiguous, and show the assembled file.

9-15.(a) Consider two count distributions of email addresses on two drives. Suppose all the counts of email addresses on the second drive are identical to the counts on the first drive. What is the cosine similarity between them assuming all the weights are 1?

(b) Suppose all the counts on the second drive are exactly half the counts on the first drive. How does this affect their cosine similarity assuming all the weights are 1?