DFRWS
DIGITAL FORENSIC RESEARCH CONFERENCE

Cross-Drive Analysis

By

Simson Garfinkel

From the proceedings of

The Digital Forensic Research Conference

DFRWS 2006 USA

Lafayette, IN (Aug 14th - 16th)

http:/dfrws.org

Digital Investigation

ELSEVIER

Forensic feature extraction and cross-drive analysis

Simson L. Garfinkel

Center for Research on Computation and Society, Harvard University, Cambridge, MA 02139, USA

Keywords:
Computer forensics
Forensic feature extraction
Cross-drive analysis
Data analysis
Information extraction

ABSTRACT

This paper introduces Forensic Feature Extraction (FFE) and Cross-Drive Analysis (CDA), two new approaches for analyzing large data sets of disk images and other forensic data. FFE uses a variety of lexigraphic techniques for extracting information from bulk data; CDA uses statistical techniques for correlating this information within a single disk image and across multiple disk images. An architecture for these techniques is presented that consists of five discrete steps: imaging, feature extraction, first-order cross-drive analysis, cross-drive correlation, and report generation. CDA was used to analyze 750 images of drives acquired on the secondary market; it automatically identified drives containing a high concentration of confidential financial records as well as clusters of drives that came from the same organization. FFE and CDA are promising techniques for prioritizing work and automatically identifying members of social networks under investigation. We believe it is likely to have other uses as well.

© 2006 DFRWS. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Most of today's forensic tools are interactive programs designed to make visible information that is hidden or hard-to-find. For example, tools such as EnCase (Guidance Software) and The Sleuth Kit (Carrier, 2005) allow an examiner to view ''deleted files'' and automatically prepare a report for use in courtroom testimony. These tools are typically designed to work with a single disk drive or image at a time.

Forensic Feature Extraction (FFE) and Cross-drive analysis (CDA) are two new approaches designed to allow an investigator to simultaneously consider information from across a corpus of many data sources, such as disk drives or solid-state storage devices. Taken together, FEE and CDA allow an investigator to rapidly identify drives of interest and to correlate or ''connect the dots'' information across multiple drives. These techniques also increase the utility of seized digital media for use in intelligence analysis.

1.1. Motivation

Today's forensic examiners have become the victims of their own success. Digital storage devices such as hard drives and flash memory are such valuable sources of information that they are now routinely seized in many investigations. As a result, examiners simply do not have the time to analyze all the media that comes across their desks.

Many organizations that engage in forensic examinations have stacks of digital media awaiting analysis. When the examiner is free, the contents of the drives are copied to a working drive. (In this paper, we will use the term ''drive'' to mean any kind of block-addressable digital media, such as disk drives or USB storage devices.) This copy or *image* is then opened or mounted using a forensic tool, after which the examiner can perform searches or manually explore the image. When finished, the image is removed from the system and the examiner proceeds to the next drive.

There are several problems with this approach:

1. *Improper prioritization.* In these days of cheap storage and fast computers, the critical resource to be optimized is the attention of the examiner or analyst. Today work is not prioritized based on the information that the drive contains.
2. *Lost opportunities for data correlation.* Because each drive is examined independently, there is no opportunity to automatically "connect the dots" on a large case involving multiple storage devices. For example, if one hard drive has an email message in the "Sent Message" mailbox and a second hard drive has that same message in an Inbox, its up to the examiner to make the connection. Such a connection would not be evident if one drive is examined on a Monday and the second is examined on the Thursday by a different examiner.
3. *Improper emphasis on document recovery.* Because today's forensic tools are based on document recovery, they have taught examiners, analysts, and customers to be primarily concerned with obtaining *documents*. Although much of the data on a typical drive cannot be reconstructed into files, this data may nevertheless be useful. The emphasis of forensic tools should be on further investigatory and evidentiary goals, not to recover files.

1.2.   Our contribution

Cross-drive analysis overcomes these problems through the use of feature extractors applied to bulk data and statistical techniques applied to a multi-drive corpus.

CDA is an outgrowth of a project in which a large number of drives were purchased on the secondary market and examined for traces of confidential information. The number of drives quickly exceeded our ability to analyze them using conventional tools. We developed a series of tools to look for credit card numbers, email addresses, and other kinds of confidential information, and then manually analyzed what these tools found. Soon we realized that these tools had general applicability beyond our immediate task and that automation could make them dramatically more powerful.

We have identified several uses for cross-drive analysis:

1. *Automatic identification of "hot" drives.* With simple statistical techniques it is possible to automatically identify drives in a large collection that are likely to be of interest, and thus should be given higher priority.
2. *Improving single drive forensic systems.* Data collected during the course of cross-drive analysis can be used to create smarter single-drive forensic tools—for example, by developing a "stop list" of information that can be safely ignored by other forensic tools.
3. *Identification of social network membership.* If several drives in a forensic repository are known to have been used by an organization under scrutiny—for example, a terrorist organization—then cross-drive analysis can be used to determine if a newly acquired piece of digital media was used by an organization who had contact with the organization in question.

4. *Unsupervised social network discovery.* Given a collection of forensic images, cross-drive analysis can be used to automatically identify organizations that were not previously known.

1.3.   Legal issues

Today's forensic investigators working on behalf of law enforcement rarely archive images from multiple investigations on a single file server. Some practitioners have argued that it is important to work on one drive at a time to avoid the inadvertent mixing of information between cases. We believe that this argument is incorrect and that it emerges from an incorrect understanding of the Federal Rules of Evidence Article 10 (US Congress, 2004), sometimes called the "best evidence rule".

Article 10 allows duplicates of original documents to be entered into evidence unless there are questions raised as the authenticity of the original or the accuracy of the copy. A law enforcement forensic lab can both implement cross-drive analysis and meet the standard set forth in Article 10 by performing the initial imaging with a hardware write-blocker and then returning the drive to the evidence locker. Should the need arise to give a copy of the drive image to opposing counsel, the drive can be retrieved from the evidence locker and imaged a second time.

A more serious legal concern is whether or not considering information from the entire disk of a suspect would conflict with the particulars of a search warrant, as some search warrants limit what can be searched by investigators. Warrants are clearly not an issue in most intelligence-related cases, and they are not an issue when permission has been given to conduct a search or when the search is not being conducted by law enforcement personnel. What is more, it may be that in the future search warrants are constructed so as to allow the *exploratory searching for terms* on the suspect's disk, but the *revealing of the search terms' context* would then require a secondary warrant. Cross-drive analysis is potentially far less invasive than other kinds of investigations because a human investigator is only exposed to information likely to yield important results. We believe that courts will welcome CDA, rather than restrict it.

1.4.   Related work

Garfinkel and Shelat developed a credit card number detector for a forensic study of 158 hard drives purchased on the secondary market (Garfinkel and Shelat, 2002). They used this detector to find drives in their corpus that contained a large number of credit card numbers for the purpose of identifying privacy violations. Garfinkel and Shelat also created a program for automatically finding email headers. This paper extends Garfinkel and Shelat's work in several important ways. First and most importantly, it introduces the idea of correlating data across multiple drives in the corpus, rather than analyzing each drive independently. Second, it expands the kinds of features considered in a multi-drive analysis; Garfinkel and Shelat only considered credit card numbers when looking for sensitive information. This paper provides the first

analysis of the pseudo-unique feature phenomena in a multi-drive forensic application. Finally, this paper is based on an analysis of a corpus that is more than five times larger than the one used by Garfinkel and Shelat.

Researchers in the field of intrusion detection have shown significant interest in using correlation as a tool for data reduction (e.g. Ning et al., 2002; Qualys, 2003; Valdes and Skinner, 2001). Such systems correlate only the events that are recognized by the IDS. Goan proposed a system for Intelligent Correlation of Evidence for automated analysis of network forensic data (Goan, 1999), but there appears to have been no follow-up work.

Existing forensic tools such as EnCase (Guidance Software) can use a database of MD5 or SHA-1 hash residues to automatically search for or suppress specific files on a hard drive that is under analysis, but not for cross-correlation.

Several tools are designed to detect or prevent the accidental leakage of private information in individual document files. For example, Computing System Innovations (CSI)'s IntelliDact can automatically scan electronic documents and even handwritten notes (once digitized) for private information such as social security numbers, bank account numbers, drives license numbers, and credit card numbers (Computing System Innovations, 2006). The software is designed to automatically detect information such as this that needs to be redacted from government documents as part of Freedom Of Information Act (FOIA) requests. Workshare Protect automatically scans documents for inadvertently privacy, intellectual property, or financial disclosures (Workshare, 2006). Neither of these products are designed to work with disk images and neither of them are designed to perform forensic analysis or data correlation.

Finally, Edelson and Krolik developed a discrete correlation function which today is widely used in astrophysics research for determining if discrete functions are correlated in the temporal domain (Edelson and Krolik, 1988). This function correlates data from two sources that are measured as different discrete instances over a long period of time. Although this function might be useful for analyzing network traffic for correlated events, it does not appear to be useful for the kind of data presented here.

1.5. Outline of this paper

Following this introduction, Section 2 presents the kinds of "features" that can be easily extracted from bulk data. Section 3 discusses how feature analysis can be applied to a single drive. Section 4 describes "first-order" applications of the cross-drive technique. Section 5 describes cross-correlation techniques applied to extracted feature sets. Section 6 describes details of our prototype cross-drive analysis system. Section 7 presents future work.

2. Forensic feature extraction

CDA is based on the identification and extraction of *pseudo-unique identifiers*, such as credit card numbers and email Message-IDs, from digital media. Once extracted, these identifiers are called "features" and are used as the basis for both single-drive analysis and multi-drive correlation.

This section discusses the principle and mathematical justification for feature extraction, presents feature extractors that we have created, and presents techniques for improving feature extraction performance.

2.1. Pseudo-unique identifiers

A pseudo-unique identifier is an identifier that has sufficient entropy such that within a given corpus it is highly unlikely that the identifier will be repeated by chance. Repetitions of pseudo-unique identifiers happen, but when they happen it is the result of a specific process, such as a file being copied from one computer to another.

An email Message-ID is a typical pseudo-unique identifier. Consider Message-ID `20060410204756.23E38908DE@spooky.sd.dreamhost.com`, which was created by the computer `spooky.sd.dreamhost.com` for an email message that was sent on April 10, 2006. The use of a time stamp, a random number and a hostname makes it very unlikely that two computers will chose the same Message-ID by accident. This is in compliance with RFC 822, which states "The uniqueness of the message identifier is guaranteed by the host which generates it" (Crocker, 1982).

Message-IDs are not unique, of course. Most Message-IDs are created for a single email message and if the same Message-ID is found on two computers, there is a good chance that an email message (or at least the Message-ID) was copied from one machine to the second.

After an email message is sent from one computer to another, both computers potentially have copies of the Message-IDs on their hard drives: those copies can be in actual files, in email message archives, in temporary files that have been deleted, or in virtual memory backing store. Multiple recipients may cause messages with the same Message-ID to travel very different paths and have different headers—even different Subject: lines, if one of the recipients is a mailing list that modifies the Subject: line. Nevertheless, the existence of the same Message-ID on two different computers strongly suggests that there was some process which transferred the identifier from the first computer to the second.

There might be alternative possible explanations for finding the same Message-ID on two different computers. For example, it is always possible that the same computer could create the same Message-ID for two different messages, although this would represent a failure of the computer's software or programming. Or two different computers could create two messages with the same Message-ID as the result of an accidental miss-configuration or an intentional spoofing attempt.

We have found that good pseudo-unique identifiers have these properties:

1. They are long enough so that collisions are unlikely to occur by chance.
2. They can be recognized using a regular expression and do not require parsing or semantic analysis.
3. They do not change over a time.

4. They can be correlated with a specific documents, people or organizations.

Not all specific identifiers in a particular class of identifiers need to be pseudo-unique. For example, the Message-ID `4231.629.XYzi-What@Other-Host` is not psuedo-unique because it appears in the text of RFC 822. As a result, any forensic tool that uses pseudo-unique identifiers needs to have a mechanism for distinguishing between identifiers that are truly pseudo-unique and those that are ubiquitous.

2.2. Feature extractors

We have built a variety of programs called *feature extractors* that can scan a disk image for pseudo-unique features and store the results in an intermediate file. Some of the feature extractors that we have built include:

- An email address extractor, which can recognize RFC822-style email addresses.
- An email Message-ID extractor.
- An email Subject: extractor.
- A Date extractor, which can extract date and time stamps in a variety of formats.
- A cookie extractor, which can identify cookies from the Set-Cookie: header in web page cache files.
- A US social security number extractor, which identifies the patterns ###-##-#### and ######### when preceded with the letters SSN and an optional colon.
- A Credit card number extractor.

Many specific features generated by these feature extractors do not meet our requirements for pseudo-uniqueness. For example, while some "Subject:" lines are certainly pseudo-unique, others are ubiquitous. Likewise, there are specific email addresses that are in Microsoft Windows DLLs and in X.509 certificates. We have developed a mathematical technique that can be used to differentiate, for example, between "Subject:" lines that are common and those that are distinctive. This technique is described in Section 4.

3. Single-drive analysis

Extracted features can be used to speed initial analysis and answer specific questions about a drive image. We have successfully used extracted features for drive image attribution and to build a tool that scans disks to report the likely existence of information that should have been destroyed under Fair and Accurate Credit Transactions Act (The fair and accurate credit transactions act of 2003, 2003).

3.1. Drive attribution

There are many circumstances where an analyst might encounter a hard drive and wish to determine to whom that drive previously belonged. For example, the drive might have been purchased on eBay and the analyst might be attempting to return it to its previous owner. Alternatively, the image might be one of several thousand obtained by spyware or another agent running on a target machine and the analyst might wish to determine if the subject belongs to a person or organization of interest. In either case, the analyst would like to have a tool so that a rapid ownership determination can be made without the need to painstakingly look for documents on the disk and then attempt to determine their pedigree and author.

We have found that a powerful technique for making this determination is to create a histogram of the email addresses on the drive (as returned by the email address feature extractor). In many cases the most common email address on the disk image is the address of the primary user of the computer from which the drive was extracted—provided that the primary user made use of email.

The most common email address on the drive is usually the email address of the drive's primary user because that person's email address appears in both the *from*: and in the *to*: or *cc*: fields of many email messages that remain on the computer's disk drive. In general, there are roughly twice as many email addresses belonging to the primary user as any other user. In our experience this is true both for users of email clients such as Outlook Express as well as for users of webmail systems such as Hotmail.

Table 1 shows a histogram of the top 15 email addresses found on Drive #51 in our collection. The first name on the list, ALICE@DOMAIN1.com, appears more than twice as much as any other name and almost certainly represents the primary use of the machine. Additional information can be readily inferred from this list. For example, the large number of email messages from JobInfo@alumni-gsb.stanford.edu strongly implies that ALICE was graduate of Stanford University's Graduate School of Business.

(Please note that all of the email addresses and many of the domain names that appear in this paper have been replaced with anonymized names in ALL CAPS in order protect the identity of the original user. In many cases the email addresses that we have found on these disk drives yield a single individual when they are typed into

Table 1 – The top 15 email addresses found on Drive #51, with the frequency of each email address	
Extracted email addresses	Count on Drive #51
ALICE@DOMAIN1.com	8133
BOB@DOMAIN1.com	3504
ALICE@mail.adhost.com	2956
JobInfo@alumni-gsb.stanford.edu	2108
CLARE@aol.com	1579
DON317@earthlink.net	1206
ERIC@DOMAIN1.com	1118
GABBY10@aol.com	1030
HAROLD@HAROLD.com	989
ISHMAEL@JACK.wolfe.net	960
KIM@prodigy.net	947
ISHMAEL-list@rcia.com	845
JACK@nwlink.com	802
LEN@wolfenet.com	790
natcom-list@rcia.com	763

Names in all caps have been used to anonymize names or domains that contain personally-identifiable information.

Table 2 – The top 15 email addresses found on Drive #80, with the frequency of each email address

Extracted email addresses	Count on Drive #80	Total drives with address
premium-server@thawte.com	117	278
server-certs@thawte.com	104	278
CPS-requests@verisign.com	61	286
personal-premium@thawte.com	44	253
personal-basic@thawte.com	42	250
personal-freemail@thawte.com	40	250
info@netscape.com	36	58
ANGIE@ALPHA.com	32	1
BARRY@BETA.com	23	1
CHARLES@GAMMA.com	21	1
DAVE.HALL@DELTA.com	21	1
DAPHNE@UNIFORM.com	20	1
ELLY@LIMA.com	18	1
FRANK@ECHO.com	16	1
HUGH@LIMA.com	16	1
IGGY@LIMA.com	16	1
GRETTA@XYZZY.com	15	1
VISTA@SNARF.com	15	1

The second column indicates the number of times that the email address was found on Drive #80, while the third column is the number of drives in the 750-image corpus on which each email address was seen.

an Internet search engine such as Google. Please also note that the drive numbers presented in this paper are based on accessioned drives, not captured images. Our corpus of 750 drive images comes from a larger collection of 1005 disk drives.)

The email histogram technique works surprisingly well even when the drive in question has not been used extensively for email. For example, Table 2 shows email addresses that were found on Drive #80, a disk that contained 1247 credit card numbers. Although the most common email addresses are from digital certificates issued by Thawte and VeriSign, if these are suppressed using techniques that will be discussed in the next section, it is possible to identify a specific email address (ANGIE@ALPHA.com) which appears to have been the primary computer user. Manual analysis of the drive revealed that the companies ALPHA.com, BETA.com, GAMMA.com, DELTA.com, UNIFORM.com and SNARF.com all make the same kind of software—and that XYZZY.COM is a personal website for an individual who uses this software and displays it on his website. It appears that Drive #80 was used to process credit cards for software that was sold by this company. This is an intelligence datum which could have been discovered through a lengthy manual examination of the drive, but which was made readily apparent through the email histogram.

4. First-order cross-drive analysis

Cross-drive analysis is the term that we have coined to describe forensic analysis of a data set that spans multiple drives. The fundamental theory of cross-drive analysis is data gleaned from multiple drives can improve the forensic

analysis of a drive in question both in the case when the multiple drives are related to the drive in question and in the case when they are not. Our architecture for cross-drive analysis uses extracted features (described above) both the make cross-drive analysis more efficient, and to focus the analysis on features that are relevant to today's forensic examinations.

This paper defines two forms of CDA: *first order*, in which the results of a feature extractor are compared across multiple drives, an $O(n)$ operation; and *second order*, where the results are correlated, an $O(n^2)$ operation.

4.1. CDA stop lists

A simple and straightforward application of CDA is to create stop lists of features that can be safely ignored in most forensic investigations because the features are ubiquitous.

For example, the first six email addresses in Table 2 are widespread on disk images today because they are present in X.509 root certificates that is distributed with many popular web browsers. Because these addresses are so widespread, they can be automatically suppressed from any list of email addresses that are displayed by forensic tools or used in further analysis. Table 3 shows the 15 email addresses that are on the largest number of drives in our corpus.

To be sure, there may be times that even ubiquitous information may be useful for an analytic process. For example, if a subject being sought is known to have used a specific version of Mozilla Firefox, then it would not make sense to suppress email addresses from certificates that were part of the Firefox distribution: to the contrary, such features could be used as a positive selection criteria in an attempt to narrow down drives that might have belong to the subject. Such a search

Table 3 – The email addresses that are observed on the largest number of drives in our 750-image corpus

Extracted email address	Drives with address	Total count in corpus
CPS-requests@verisign.com	286	64,424
server-certs@thawte.com	278	32,873
premium-server@thawte.com	278	31,141
Mouse.Exe@Mouse.com	262	493
LMouse.Exe@LMouse.com	262	493
personal-premium@thawte.com	253	14,660
personal-freemail@thawte.com	250	14,843
personal-basic@thawte.com	250	14,290
inet@microsoft.com	244	31,456
mazrob@panix.com	221	3265
java-security@java.sun.com	200	1200
java-io@java.sun.com	198	413
someone@microsoft.com	195	6193
bugs@java.sun.com	192	351
ca@digsigtrust.com	173	36,800
name@company.com	169	1763

These email addresses (and many others) can be automatically suppressed by forensic tools because they are part of the operating system and, therefore, not likely to be related to a case under investigation. (The email address `mazrob@panix.com` is present in the Windows system file `clickerx.wav` and appears to be the email address of the authors of the ''Close Program'' sound for the Windows 95 Utopia Sound Scheme.)

represents a very specific application which can easily be handled by simply turning off the stop list: this application shows why the stop list should be used to suppress *output*, rather than for suppressing *collection*.

4.2. Hot drive identification

If the features extracted from the disk images are generically of interest to the investigator, then the investigator's work can be easily prioritized by concentrating on the drives that have the largest number of these features. We call this kind of prioritization "hot drive identification".

For example, the Fair and Accurate Credit Transactions Act (The fair and accurate credit transactions act of 2003, 2003) of 2003 (FACT-ACT) requires that US corporations disposing of electronic media to purge the media of "consumer information". The US Federal Trade Commission's Final Rule implementing the rule defines consumer information as "including, but not limited to, a social security number, driver's license number, phone number, physical address, and email address" (Federal Trade Commission, 2004).

Because we have feature extractors that can recognize social security numbers, email addresses and other "consumer information", we can automatically identify violations of the FACT-ACT. Work can be automatically prioritized by querying the database for the drives with the largest number of features that correspond to "consumer information".

Our social security number extractor was able to find identified social security numbers in 48 of the 750 disk images. Of these nine contained SSNs that appeared to be test data (e.g. 555-55-5555 and 666-66-6666). Eliminating these test SSNs left 39 disks that had SSNs which represented privacy violations. One of these, Drive #959, had 260 unique SSNs and appears to contain consumer credit applications. Table 4 shows the disks with the most SSNs. An organization charged with policing for violations of the FACT-ACT could use this list to prioritize its work.

As a second example of this "hot drive" technique, we computed histograms of the extracted email addresses for our corpus of 750 images. We found 13 drives (#339, #340, #342, #343, #345, #356, #348–#351, #354, #356 and #357), each approximately 1 GB in size, that all had between 710,000 and 765,000 unique email addresses and between 2.4 million and 2.7 million email addresses in total. These drives, which we will

call "Lot SP", were obtained as the result of a single purchase brokered through eBay. (Overall, the 750 drive images in the corpus represent approximately 75 lots.) The drives in Lot SP appears to come from an organization that was involved in sending bulk email: for example, many of the email addresses on these drives appear in alphabetical order, sorted by domain name to allow for efficient use of SMTP connections, many clearly do not belong to individuals (e.g. test.agent1@some domain followed by test.again@somedomain) and many appear to have been scraped from web pages.

Not only does the sale of these drives on eBay possibly represent a violation of the FACT-ACT, data on the drives may also indicate that additional laws restricting the sending of bulk email have been violated. Although we did not set out to find individuals or organizations engaged in such practices, these hot drives were readily apparent.

5. Second-order cross-drive analysis

Section 4 explored a variety of first-order cross-drive analysis. This section explores second-order techniques that are based on cross-correlations the data on multiple drives. Put another way, in Section 4 we explored techniques for automatically selecting drives that had the largest number of email addresses and other features. In this section we explore a different question: *which are the drives in the corpus that have the largest number of features in common?* This question can be answered using multi-drive correlation of discrete features.

We created a *Multi-Drive Correlator* (MDC)—a program that reads multiple feature files and produce report containing, for each feature, a list containing the number of drives on which that feature was seen, the total number of times that feature was seen on all drives, and a list of the drives on which that feature occurs. Mathematically, the MDC is a function whose input is a set of drive images in a feature to be correlated, and whose output is a list of (*feature, drive—list*) touples.

5.1. Email address multi-drive correlation

Applying the MDC to the email feature files, we learned that the corpus contained 6,653,396 unique email addresses. Because so many email addresses were found on the 13 drives of "Lot SP", these drives were suppressed and a second MDC was calculated: without Lot SP, there were only 331,186 unique email addresses in the corpus. A histogram analysis of both correlations appears in Table 5. The first line of the table shows how many unique email addresses were found on a single drive, the second line shows how many unique email addresses were found on just two drives, and so on. This table implies that the number of email addresses in common between drive images seems to follow a power-law distribution. We have found such distributions to be common when performing MDC analyses.

5.2. Scoring the correlation

Once the correlation list is produced, it is desirable to produce a report of the drives that are most highly correlated. We have experimented with three weighting functions for scoring the correlation between each pair of drives.

Table 4 – Drives images with the most extracted Social Security Numbers (SSNs), after obvious test data have been suppressed

Drive	Unique SSNs	Total SSNs
Drive #959	260	447
Drive #974	178	674
Drive #696	33	872
Drive #969	33	33
Drive #690	8	14
Drive #680	2	4

"Unique SSNs" is the number of individual SSNs that were found, while "Total SSNs" is the total number of SSNs that were present, including duplicates.

Table 5 – The total number of email addresses found on a single drive, on a pair of drives, and so on

# Of drives with common email addresses	Number email addresses in common	
	Entire corpus	Without "Lot SP"
1	4,903,909	331,186
2	1,145,507	15,909
3	209,774	2914
4	108,909	1623
5	59,550	2086
6	41,816	536
7	31,767	437
8	23,881	309
9	20,337	164
10	18,269	81
11	17,134	66
12	18,427	61
13	53,209	56
14	248	43
…	…	…
250	2	2
253	1	1
262	2	2
278	2	2
286	1	1
Total email addresses	6,653,396	356,037

The middle column shows the number of email addresses found on all drives in the corpus, while the right column shows the number of email addresses found on all of the drives in the corpus with the exception of those drives that were in "Lot SP".

Let:

D = # of drives
F = # of extracted features
$d_0 \dots d_D$ = Drives in corpus
$f_0 \dots f_F$ = Extracted features
$FP(f_n, d_n) = \begin{cases} 0 & f_n \text{ not present on } d_n \\ 1 & f_n \text{ present on } d_n \end{cases}$

A simple scoring function is to add up the number of features that two drives have in common:

$$S_1(d_1, d_2) = \sum_{n=0}^{F} FP(f_n, d_1) FP(f_n, d_2)$$

A more sophisticated weighting function discounts features by the number of drives on which they appear, which makes correlations resulting from pseudo-unique features more important than correlations based on ubiquitous features:

$$DC(f) = \sum_{n=0}^{D} FP(f, d_n) = \text{\# of drives with feature } f$$

$$S_2(d_1, d_2) = \sum_{n=0}^{F} \frac{FP(f_n, d_1) FP(f_n, d_2)}{DC(f_n)}$$

Perhaps rare features that are present in high concentrates on drives d_1 and/or d_2 should increase the weight. This would increase the score between a computer user who had exchange a lot of email with a known terrorist when compared with an individual who has only exchanged one or two emails with the terrorist:

$FC(f, d)$ = count of feature f on drive d

$$S_3(d_1, d_2) = \sum_{n=0}^{F} \frac{FC(f_n, d_1) FC(f_n, d_2)}{DC(f_n)}$$

We are in the process of evaluating these three weighting functions; our initial findings are reported below.

5.3. A scored SSN correlation

We performed an MDC using the extracted social security numbers. After removing spaces and dashes from the recognized SSNs, we found only five SSNs are were present on more than one drive (Table 6). Although a total of 571 SSNs were found in the 750-drive corpus, only five SSNs were found on more than one drive. Of these, three were test SSNs and two appear to be valid SSNs which we shall call SSN_1 and SSN_2 for the purpose of this paper:

- SSN_1 was found on three drives: Drive #342, #343 and #356. In each case the SSNs appeared in unstructured text. Before the SSN was a date of birth of April 27, 19XX. After the SSN was the notation "Thanks, Laurie". All of these drives were purchased as part of Lot 34 and all appear to have come from the same organization.
- SSN_2 was found on two drives: Drive #350 and #355. In both images the SSN is preceded with the string "great grandchildren" and followed by the string "I used to". Because the SSN appears at different locations in the two disk images, we believe that the information was copied from one drive to the second in the course of normal computer operations. Both drives are SCSI Seagate ST19171W drives with a SUN9.0G firmware and of exactly the same size.

Reviewing Table 7, function S_3 gave drive pair (612, 690) the highest weight. This makes sense, since these two drives together had eight copies of the SSN "555555555". The fact that this is a test social security number and not a real one is ironic but, ultimately, irrelevant. S_3's real failure is that it does not correlate the three drives with SSN_1 as strongly as the nine drives with the SSN "666666666".

Interestingly, due to a clerical error at the time of imaging, the data for Drive #355 was originally labeled as coming from Drive #357, which is from a different lot. After the correlation

Table 6 – A multi-drive correlation of SSNs

SSN	Found on drives	Total found
666666666	313, 427, 429, 430, 612, 627, 744, 770, 808	11
123456789	328, 343, 345, 350, 351, 700	8
SSN_1	342, 343, 356	3
555555555	612, 690	8
SSN_1	350, 357	2

Unlike Table 4, test data have not been suppressed. The numbers SSN_1 and SSN_2 have been anonymized because they represent actual SSNs belonging to individuals.

Table 7 – The result of the three scoring functions presented in Section 5.2 applied to some of the drive pairs in Table 6, sorted by S_3 scores

Drive pair	S_1	S_2	S_3
(612, 690)	1.000	0.500	8.000
(350, 700)	1.000	0.167	0.667
(350, 357)	1.000	0.500	0.500
(612, 744)	1.000	0.111	0.444
(351, 700)	1.000	0.167	0.333
(345, 350)	1.000	0.167	0.333
(342, 356)	1.000	0.333	0.333
(328, 700)	1.000	0.167	0.333
(342, 343)	1.000	0.333	0.333
(328, 350)	1.000	0.167	0.333
(343, 356)	1.000	0.333	0.333
(343, 700)	1.000	0.167	0.333
(343, 350)	1.000	0.167	0.333
…	…	…	…

match was noted, we carefully examined the metadata associated with the drives and the actual drives to verify the cross-lot correlation and discovered our error. We were able to determine the ground truth of Drive #355 because the drive was physically labeled with both its lot number and drive number, and because our disk imaging program records both the bytes read from the drive and the drive's serial number in a single file. This example shows both the importance of recording data with metadata, a point made by Garfinkel et al. (Garfinkel et al., in press), and the power of the cross-drive correlation technique for identifying drives from the same organization. It also shows how CDA can be used for social network analysis: in this case, the network that was discovered were the network that contained drives (342, 343, 356) and the network that contained drives (350, 357).

5.4. A credit card number MDC

A total of 5,796,217 strings of 14, 15 and 16-digit numbers in the 750-drive corpus passed our first CCN test, while only 159,419 passed all four tests (Fig. 1). We applied the multi-drive correlator to both collections and then computed the drive-pair weights for each correlation result. We had previously identified three pairs of drives in first set of 250 disk images as being highly correlated: one pair (171, 172) was correlated because of actual credit card numbers, while

1. The string is a sequence of 14–16 digits with either no spaces or broken up by spaces or dashes in the manner that credit card numbers are typically displayed.

2. No single digit is repeated more than 7 times, and no pairs of digits are repeated more than 5 times.

3. The first 4 digits belongs to financial institution that is known to issue credit cards, and the length of the string without spaces is consistent with the particular financial organization.

4. The sequence of digits follows the credit card number validation algorithm.

Fig. 1 – The four tests used by the credit card number feature extractor.

two other pairs (74, 77) and (179, 206), had been correlated the basis of string sequences that passed the CCN-identifier test, but which actually were not. Each of these pairs was apparently correlated because both halves of the pair contained the same fragments of a file that had the false-positives.

Because of the large number of drives with CCNs in our data set, the remainder of this section looks at just a few pairs that we have considered. Table 8 notes the maximum score for all drive pairs using both corpa as well as the score of several notable drive pairs discussed below:

- *Drives #74 and #77*. These two drives were part of a lot purchased from a single reseller in the Pacific Northwest. Manual inspection of the information on the drives had previously revealed that four of them had come from the same community college. The cross-correlation found 25 unique 15 and 16-digit numeric strings that were recognized as CCNs by the CCN feature extractor that were common to these drives and only to these drives, but visual inspection revealed that they were not actually CCNs, but instead false-positives of the CCN detector.
- *Drives #171 and #172*. The first-order analysis of our corpus identified Drive #172 as being of interest because of the large number of CCNs that it contained—31,348 CCNs, of which 11,609 (37%) were unique. This drive was later manually identified as being an Oracle database drive that had been used to hold patient billing records by a medical center in Florida.

The cross-drive analysis revealed that this drive had 13 unique CCNs in common with Drive #171. Unlike the previous example, these identifiers appear to be actual CCNs. Subsequent analysis of Drive #171 revealed that this drive contained 346 CCNs, of which 81 (23%) were unique. Also found on Drive #171 was C source code. It appears likely that this drive was used by the medical center's programmers for their development system, and that the programmers tested their system with actual patient data.

- *Drives #339 through #356*. These drives were all purchased from a dealer in New York, NY. Manual inspection reveals that many of these drives were used by a travel agency; many contained names, credit card numbers, ticket

Table 8 – Results of the scored multi-drive correlation applied to the corpus of CCNs that passed the first test in Fig. 1, and those that pass all of the tests

	5,796,217 CCN corpus			159,419 CCN corpus		
	S_1	S_2	S_3	S_1	S_2	S_3
Max score	6817	3047	7,453,650	236	61	16,459
(74, 77)	748	319	394	18	9	9
(171, 172)	1487	742	7,456,650	7	3.5	3736
(345, 350)	671	129	2608	203	52	885
(350, 356)	825	175	1863	236	61	556
(695, 698)	334	13	3,861,670	1	0.055	0.055
(716, 718)	6817	3047	20,638	38	14	14
(814, 820)	571	122	997,384	3	1	1

numbers, itineraries, and email messages to clients. A cluster analysis, which will be described in a future paper, shows that all of these drives are highly correlated using many different weights. A representative drive pair is reported in Table 8.

- *Drives #716 and #718*. These two drives were both part of Lot 70, a collection of four drives from a dealer in Union City, CA. We have not done further analysis to understand why these drives are correlated.
- *Drives #814 and #820*. These two drives were part of Lot 78, a collection of 15 drives purchased from a dealer in Stamford, CT. As with the previous drives, we have not yet determined why these drives are correlated.

6.   Implementation

We have designed an end-to-end architecture for cross-drive analysis that accessions and images data from disk drives and other digital storage media obtained on the secondary market, stores intermediate results in feature files and a database, builds intermediate cross-correlation tables, and supports an interactive multi-user interface for database exploration. Data flows through the system in a series of steps:

1. Disks collected on the secondary market are imaged onto into a single AFF file. (AFF is the Advanced Forensic Format, a file format for disk images that contains all of the data accession information, such as the drive's manufacturer and serial number, as well as the disk contents. AFF also has the ability to distinguish sectors that cannot be read from sectors that are properly cleared. As an added benefit, AFF stores the disk image as a series of compressed segments, dramatically minimizing the amount of server space consumed by the image while still allowing the data within the image to be randomly accessed that contains both the drive's data and metadata such as the drive's serial and model numbers (Garfinkel et al., in press).
2. The `afxml` program is used to extract drive metadata from the AFF file and build an entry in the SQL database.
3. Strings are extracted with an AFF-aware program in three passes, one for 8-bit characters, one for 16-bit characters in lsb format, and one for 16-bit characters in msb format.
4. Feature extractors run over the string files and write their results to feature files.
5. Extracted features from newly-ingested drives are run against a *watch list*; hits are reported to the human operator.
6. The feature files are read by indexers, which build indexes in the SQL server of the identified features.
7. A multi-drive correlation is run to see if the newly accessioned drive contained features in common with any drives that are on a *drive watch list*.
8. A user interface allows multiple analysts to simultaneously interact with the database, to schedule new correlations to be run in a batch mode, or to view individual sectors or recovered files from the drive images that are stored on the file server.

Our proof-of-concept system has advanced to the point that we could use it to prepare this paper and run a few additional experiments.

6.1.   Extractor implementation

Our feature extractors are based on regular expressions compiled with flex (Paxson, 1995). Additional rules are implemented in C++.

Although it is possible to run the scanners directly on raw ("dd") disk images, we have found that an improved technique is to first preprocess the disk images with the `strings` program that is part of the Free Software Foundation's `binutils` distribution. Three passes are made with `strings`, extracting 8-bit-byte, 16-bit bigendian, and 16-bit littleendian codings. We then run the scanners on the resulting files. In this manner, the amount of data that the feature extractors need to examine is reduced, while the amount of features that can be extracted is actually increased (since an extractor written to recognize 8-bit features can now find 8-bit features that are coded in 16-bit character sets).

The results of each extractor are saved in a *feature file*. Each line of the file consists of the feature that was detected, the context in the file before and after the feature, and the offset of the feature in the disk image. Both the context and the position information can be used by other tools—for example, by an interactive tool that allows an analyst to view the region in the file system where the feature was detected. An example of a feature file appears in Fig. 2.

6.2.   Correlator implementation

Our initial MDC was written in python; although python is a lovely language for prototyping, we found it to be relatively slow and memory-intensive for this work: performing the MDC of the email addresses resulted in a python process that slowly expanded to consume more than 3.5 GB of memory and did not complete its task after 24 h of computation due to excessive paging. Rewriting the MDC in a mixture of C and C++ resulted in a fast correlator that consumed less than 600 MB of memory; correlations of our 750-drive corpus typically take between 10 min and 2 h on a 1.8 GHz AMD64. The MDC uses a hash table based Goldfoot's "Simple Hash" implementation (Goldfoot, 2006). In exchange for speed, this implementation does not include features such as data generalization or automatic re-hashing: hash tables must be declared to be a particular size when they are first created.

7.   Future work

Cross-drive analysis shows significant promise as a technique for improving the automation of forensic tools and for intelligence analysis. Further work in this also will require progress on four fronts:

First, we need to do increase our understanding of the multi-drive correlation, and in particular techniques that

```
EMAIL|n.com; by E-mail at |CPS-requests@verisign.com|; or.by mail at Veri  (pos=3581922)
COOKIE|s","CachePrefix",2,"|Cookie:"|.HKLM,"Software\Micr  (pos=3849059)
EMAIL|n.com; by E-mail at |CPS-requests@verisign.com|; or.by mail at Veri  (pos=6982915)
EMAIL|emium Server CA1(0&.|premium-server@thawte.com|0.960801000000Z.2012  (pos=9441431)
EMAIL|emium Server CA1(0&.|premium-server@thawte.com|0.H5:R.x`^^n7c"w6~.W  (pos=9441602)
SUBJECT|: .Sent: .To: .Cc: .|Subject: |.Importance: .Sensit  (pos=35418278)
SUBJECT|sation: .Keywords: .|Subject: |.Importance: .Sensit  (pos=35423128)
COOKIE|txt.URL .TgvH.z\gvH.|Cookie:SELJEJN@iwon.com/|.SELJEJN@iwon[1].txt  (pos=57277759)
COOKIE|jn@iwon[1].txt.URL .|Cookie:SELJEJN@virtupay.net/|.SELJEJN@virtupay  (pos=57277809)
```

Fig. 2 – Example lines of a feature file produced by a feature extractor. The username in the cookies in the last two lines have been anonymized.

can be used to more accurately score the relationship between drive pairs and to cluster drives.

Second, we need to improve our facility at working with the large data sets required to do cross-drive analysis. We believe that there are many opportunities to improve performance, including the use of machines with larger main memories; developing algorithms designed to run on clusters; and the use of more efficient algorithms.

Third, we need better feature extractors. For example, we plan to extend the cookie extractor to extract cookies from cookie jars. Additional specificity will be achieved by preprocessing the disk images using a forensic tool such as The Sleuth Kit to extract all data files from the disk image and then using format-specific feature extractors. In the future we also hope to use language-aware systems such as the Rosette Linguistics Platform (Basis Technology, 2006). We also hope to create a system that performs correlations based on cryptographic hashes of individual sectors in the disk images. (An interesting property of most modern file systems is that files larger than 4K are invariably stored with their first bytes block-aligned. Thus, any search for the MD5s of the file's "sectors" will appear on the hard drive, even if the file system format is not understood. It should be possible to use as features the hashes of all of the sectors of a disk drive.)

Finally, we need to develop tools that can make this technique useful to forensic workers and intelligence analysts. Although our preference is the creation of automated tools, at first it is likely that it will be easier to create interactive tools that leverage pre-computed feature indexes.

Acknowledgments

REFERENCES

Basis Technology. Rosette linguistics platform, <http://www.basistech.com/products/>; 2006.

Carrier Brian. The Sleuth Kit & Autopsy: forensics tools for Linux and other Unixes, <http://www.sleuthkit.org/>; 2005.

Computing System Innovations. Intellidact: CSI technology for automated redaction and indexing, <http://www.csisoft.com/applications/intellidact.php>; 2006.

Crocker D. RFC 822: standard for the format of ARPA Internet text messages; August 13, 1982. See also STD0011. Obsoletes RFC0733. Updated by RFC1123, RFC1138, RFC1148, RFC1327, RFC2156. Status: STANDARD.

Edelson RA, Krolik JH. The discrete correlation function: a new method for analyzing unevenly sampled variable data. The Astrophysics Journal 15 October 1988;333:646–59.

Federal Trade Commission. Disposal of consumer report information and records, final rule, <http://www.ftc.gov/os/2004/11/041118disposalfrn.pdf>; November 2004.

Garfinkel Simson, Shelat Abhi. Remembrance of data passed. IEEE Security and Privacy Magazine; January 2002.

Garfinkel Simson L, Malan David J, Dubec Karl-Alexander, Stevens Christopher C, Pham Cecile. Disk imaging with the advanced forensic format, library and tools. In: Research advances in digital forensics (second annual IFIP WG 11.9 international conference on digital forensics). Springer, in press.

Goan Terrance. A cop on the beat: collecting and appraising intrusion evidence. Communications of the ACM 1999;42(7). ISSN: 0001-0782:46–52.

Goldfoot Josh. The computer language shootout, <http://shootout.alioth.debian.org/>; 2006.

Guidance Software, Inc. EnCase Forensic, <http://www.guidancesoftware.com/products/ef_index.asp>.

Ning Peng, Cui Yun, Reeves Douglas S. Constructing attack scenarios through correlation of intrusion alerts. In: CCS'02 proceedings of the ninth ACM conference on computer and communications security. New York, NY, USA: ACM Press, ISBN 1-58113-612-9; 2002. p. 245–54.

Paxson Vern. Flex, version 2.5, a fast scanner generator. 2.5 edition. The Free Software Foundation, <http://www.gnu.org/software/flex/manual/>; March 1995.

Qualys. Qualys IDS correlation daemon, <http://quidscor.soruceforge.net/>; October 2003.

The fair and accurate credit transactions act of 2003; 2003 [Public Law 108–159, 117 Stat. 1952].

US Congress. Federal rules of evidence, <http://www.law.cornell.edu/rules/fre/>; 2004.

Valdes Alfonso, Skinner Keith. Probabilistic alert correlation. In: Proceedings of the fourth international symposium on recent advances in intrusion detection (RAID), vol. 2212 [Lecture notes in Computer Science]; 2001.

Workshare. Workshare unveils the workshare protect enterprise suite to cure information leakage without crippling business, <http://www.workshare.com/company/news/pressreleases/prelease_54.aspx>; 2006.

Simson L. Garfinkel is a postdoctoral fellow at the Center for Research on Computation at Society at Harvard University, and a research affiliate at the Computer Science and Artificial Intelligence Laboratory at MIT. He is also a consulting scientist at Basis Technology Corp., which develops software for extracting meaningful intelligence from unstructured text, and a founder of Sandstorm Enterprises, a computer security firm that develops advanced computer forensic tools used by businesses and governments to audit their systems. Dr. Garfinkel has research interests in computer forensics, the emerging field of usability and security, and in personal information management.