Forensic Feature Extraction and Cross-Drive Analysis

Simson L. Garfinkel
Center for Research on Computation and Society Harvard University
1:15pm, Tuesday, August 15, 2006
Today’s forensic tools are designed for one drive at a time.

Primary Goals: Search and Recovery.

Interactive user interface.

Usage scenarios:

- Recovery of “deleted” files.
- Child porn scanning.
- Trial preparation.
Today’s tools choke when confronted with hundreds or thousands of drives.

Which drives were used by my target?
Do any drives belong to the target’s associates?
Who is talking to who?
Where should I start?

Police departments and intelligence agencies have thousands of drives...
Additional problems with today’s tools

• Improper prioritization
  Letting priority be determined by the statute of limitations.

• Lost opportunities for data correlation
  Was a message on hard drive X sent to hard drive Y?

• Emphasis on document recovery rather than in furthering the investigation.
Correlating data *between* drives is an untapped opportunity.

How large is my target’s reach?
Who is in the organization?

Captured drives are an ideal social network analysis.
This talk introduces Cross Drive Analysis

Large scale forensics problem

Architecture

Feature extraction

First order analysis

Second order analysis
Forensic Feature Extraction and Cross-Drive Analysis

1. Get a lot of drives
2. Image to a big disk
3. Extract the Features
4. Apply statistics and correlation
Uses of Cross-Drive Analysis

1. Automatic identification of hot drives
2. Improvements to single-drive systems
3. Identification of social network membership
4. Unsupervised social network discovery

Related Work:

• Garfinkel & Shelat, 158 drives, 2002
• FTK 2.0 — indexing multiple drives
• IntelliDact and Workshare Protect scan for confidential information
Feature extractors find *pseudo-unique* features

Pseudo-Unique characteristics:

- Long enough so collisions by chance are unlikely.
- Recognizable with regular expressions.
- Persistent over time.
- Correlated with specific documents, people or organizations.

Typical Features:

- email addresses
- Message-IDs
- Subject: lines
- Cookies
- US Social Security Numbers
- Credit card numbers
- Hash codes of drive sectors
Example: The Credit Card Number Detector.

The CCN detector scans bulk data for ASCII patterns that look like credit card numbers.

- CCNs are found in certain typographical patterns. (e.g. XXXX-XXXX-XXXX-XXXX
or XXXX XXXX XXXX XXXX
or XXXXXXXXXXXXXXXXXXXX )

- CCNs are issued with well-known prefixes.

- CCNs follow the Credit Card Validation algorithm.

- Certain numeric patterns are unlikely. (e.g. 4454-4766-7667-6672)
CCN detector: written in flex and C++

Scan of Drive #105: (642MB)

<table>
<thead>
<tr>
<th>Test</th>
<th># pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>typographic pattern</td>
<td>3857</td>
</tr>
<tr>
<td>known prefixes</td>
<td>90</td>
</tr>
<tr>
<td>CCV1</td>
<td>43</td>
</tr>
<tr>
<td>numeric histogram</td>
<td>38</td>
</tr>
</tbody>
</table>

Sample output:

'value CHASE NA | 5422-4128-3008-3685 | pos=13152133
'value DISCOVER | 6011-0052-8056-4504 | pos=13152440
'value GE CARD | 4055-9000-0378-1959 | pos=13152589
'value BANK ONE | 4332-2213-0038-0832 | pos=13152740
'value NORWEST | 4829-0000-4102-9233 | pos=13153182
'value SNB CARD | 5419-7213-0101-3624 | pos=13153332
Even with the tests, there are occasional false positives.

CCN scan of Drive #115: (772MB)

<table>
<thead>
<tr>
<th>Test</th>
<th># pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>pattern</td>
<td>9196</td>
</tr>
<tr>
<td>known prefixes</td>
<td>898</td>
</tr>
<tr>
<td>CCV1</td>
<td>29</td>
</tr>
<tr>
<td>patterns</td>
<td>27</td>
</tr>
<tr>
<td>histogram</td>
<td>13</td>
</tr>
</tbody>
</table>

.................@: |444444866666108| :<@<744444:@@<<44 pos=82473275
.................#"&’’|445447667667667| ..050014’’4"1"&’ pos=86493675
.............221267241667&|454676676654450|&566746566726322 pos=86507818
3..30210212676677..|30232676630232| .1...........001.01 pos=86516059
"&##’’&41’&’645445&|454445672676632| .3.............0.. pos=86523223
.................".##"##’’|445467667227023| ............366 pos=87540819
D#9?.32400. ,+14%B|499745255278101| *02)46+; <17756669 pos=118912826
.GGJJB...>.JJGG...G|3534554333511116| .................6 pos=197711868
%.....}{})})}}}}........|44444322233345|......}{})})}}}}}} pos=228610295
%6"!) .&*%, ,%-0)07 .|373484553420378|<67<038+.5(+0+.3 pos=638491849
%6"!) .&*%, ,%-0)07 .|373484553420378|<67<038+.5(+0+.3 pos=645913801
CDA Prototype System

750 images
1.5TB data compressed.
Many different organizations.
Single-drive feature application: drive attribution.

Drive #51: Top email addresses (sanitized)

<table>
<thead>
<tr>
<th>Address</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:ALICE@DOMAIN1.com">ALICE@DOMAIN1.com</a></td>
<td>8133</td>
</tr>
<tr>
<td><a href="mailto:BOB@DOMAIN1.com">BOB@DOMAIN1.com</a></td>
<td>3504</td>
</tr>
<tr>
<td><a href="mailto:ALICE@mail.adhost.com">ALICE@mail.adhost.com</a></td>
<td>2956</td>
</tr>
<tr>
<td><a href="mailto:JobInfo@alumni-gsb.stanford.edu">JobInfo@alumni-gsb.stanford.edu</a></td>
<td>2108</td>
</tr>
<tr>
<td><a href="mailto:CLARE@aol.com">CLARE@aol.com</a></td>
<td>1579</td>
</tr>
<tr>
<td><a href="mailto:DON317@earthlink.net">DON317@earthlink.net</a></td>
<td>1206</td>
</tr>
<tr>
<td><a href="mailto:ERIC@DOMAIN1.com">ERIC@DOMAIN1.com</a></td>
<td>1118</td>
</tr>
<tr>
<td><a href="mailto:GABBY10@aol.com">GABBY10@aol.com</a></td>
<td>1030</td>
</tr>
<tr>
<td><a href="mailto:HAROLD@HAROLD.com">HAROLD@HAROLD.com</a></td>
<td>989</td>
</tr>
<tr>
<td><a href="mailto:ISHMAEL@JACK.wolfe.net">ISHMAEL@JACK.wolfe.net</a></td>
<td>960</td>
</tr>
<tr>
<td><a href="mailto:KIM@prodigy.net">KIM@prodigy.net</a></td>
<td>947</td>
</tr>
<tr>
<td><a href="mailto:ISHMAEL-list@rcia.com">ISHMAEL-list@rcia.com</a></td>
<td>845</td>
</tr>
<tr>
<td><a href="mailto:JACK@nwlink.com">JACK@nwlink.com</a></td>
<td>802</td>
</tr>
<tr>
<td><a href="mailto:LEN@wolfenet.com">LEN@wolfenet.com</a></td>
<td>790</td>
</tr>
<tr>
<td><a href="mailto:natcom-list@rcia.com">natcom-list@rcia.com</a></td>
<td>763</td>
</tr>
</tbody>
</table>

Most common email address is (usually) drive’s primary user.
Attribution histogram works even with lightly-used drives.

<table>
<thead>
<tr>
<th>Extracted Email Addresses</th>
<th>Count on Drive #80</th>
<th>Total drives with address</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:premium-server@thawte.com">premium-server@thawte.com</a></td>
<td>117</td>
<td>278</td>
</tr>
<tr>
<td><a href="mailto:server-certs@thawte.com">server-certs@thawte.com</a></td>
<td>104</td>
<td>278</td>
</tr>
<tr>
<td><a href="mailto:CPS-requests@verisign.com">CPS-requests@verisign.com</a></td>
<td>61</td>
<td>286</td>
</tr>
<tr>
<td><a href="mailto:personal-premium@thawte.com">personal-premium@thawte.com</a></td>
<td>44</td>
<td>253</td>
</tr>
<tr>
<td><a href="mailto:personal-basic@thawte.com">personal-basic@thawte.com</a></td>
<td>42</td>
<td>250</td>
</tr>
<tr>
<td><a href="mailto:personal-freemail@thawte.com">personal-freemail@thawte.com</a></td>
<td>40</td>
<td>250</td>
</tr>
<tr>
<td><a href="mailto:info@netscape.com">info@netscape.com</a></td>
<td>36</td>
<td>58</td>
</tr>
<tr>
<td><a href="mailto:ANGIE@ALPHA.com">ANGIE@ALPHA.com</a></td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:BARRY@BETA.com">BARRY@BETA.com</a></td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:CHARLES@GAMMA.com">CHARLES@GAMMA.com</a></td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:DAVE.HALL@DELTA.com">DAVE.HALL@DELTA.com</a></td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:DAPHNE@UNIFORM.com">DAPHNE@UNIFORM.com</a></td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:ELLY@LIMA.com">ELLY@LIMA.com</a></td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:FRANK@ECHO.com">FRANK@ECHO.com</a></td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:HUGH@LIMA.com">HUGH@LIMA.com</a></td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:IGGY@LIMA.com">IGGY@LIMA.com</a></td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:GRETTA@XYZZY.com">GRETTA@XYZZY.com</a></td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td><a href="mailto:VISTA@SNARF.com">VISTA@SNARF.com</a></td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

Email addresses found on ≈> 20 drives are not pseudo-unique
First Order Cross-Drive Analysis:
$O(n)$ operations on feature files

Applications:

- Automatically building stop lists
- Hot drive identification
Automatic “stop lists:”
features on many drives are not pseudo-unique.

<table>
<thead>
<tr>
<th>Extracted Email Address</th>
<th>Drives with address</th>
<th>Total count in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:CPS-requests@verisign.com">CPS-requests@verisign.com</a></td>
<td>286</td>
<td>64424</td>
</tr>
<tr>
<td><a href="mailto:server-certs@thawte.com">server-certs@thawte.com</a></td>
<td>278</td>
<td>32873</td>
</tr>
<tr>
<td><a href="mailto:premium-server@thawte.com">premium-server@thawte.com</a></td>
<td>278</td>
<td>31141</td>
</tr>
<tr>
<td><a href="mailto:Mouse.Exe@Mouse.Com">Mouse.Exe@Mouse.Com</a></td>
<td>262</td>
<td>493</td>
</tr>
<tr>
<td><a href="mailto:LMouse.Exe@LMouse.Com">LMouse.Exe@LMouse.Com</a></td>
<td>262</td>
<td>493</td>
</tr>
<tr>
<td><a href="mailto:personal-premium@thawte.com">personal-premium@thawte.com</a></td>
<td>253</td>
<td>14660</td>
</tr>
<tr>
<td><a href="mailto:personal-freemail@thawte.com">personal-freemail@thawte.com</a></td>
<td>250</td>
<td>14843</td>
</tr>
<tr>
<td><a href="mailto:personal-basic@thawte.com">personal-basic@thawte.com</a></td>
<td>250</td>
<td>14290</td>
</tr>
<tr>
<td><a href="mailto:inet@microsoft.com">inet@microsoft.com</a></td>
<td>244</td>
<td>31456</td>
</tr>
<tr>
<td><a href="mailto:mazrob@panix.com">mazrob@panix.com</a>(*)</td>
<td>221</td>
<td>3265</td>
</tr>
<tr>
<td><a href="mailto:java-security@java.sun.com">java-security@java.sun.com</a></td>
<td>200</td>
<td>1200</td>
</tr>
<tr>
<td><a href="mailto:java-io@java.sun.com">java-io@java.sun.com</a></td>
<td>198</td>
<td>413</td>
</tr>
<tr>
<td><a href="mailto:someone@microsoft.com">someone@microsoft.com</a></td>
<td>195</td>
<td>6193</td>
</tr>
<tr>
<td><a href="mailto:bugs@java.sun.com">bugs@java.sun.com</a></td>
<td>192</td>
<td>351</td>
</tr>
<tr>
<td><a href="mailto:ca@digsigtrust.com">ca@digsigtrust.com</a></td>
<td>173</td>
<td>36800</td>
</tr>
<tr>
<td><a href="mailto:name@company.com">name@company.com</a></td>
<td>169</td>
<td>1763</td>
</tr>
</tbody>
</table>

*mazrob@panix.com appears in clickerx.wav (Utopia Sound Scheme)*
A graph of # email addresses on each drive automatically identified drives used by bulk e-mailers.
Hot drive identification:
Drives with high response warrant further attention.

Only 7 drives had more than 300 credit card numbers.
Hot drive identification: Drives with high response warrant further attention.

These drives represent significant privacy violations.
First order analysis of # SSNs

<table>
<thead>
<tr>
<th>Drive</th>
<th>Unique SSNs</th>
<th>Total SSNs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive #959</td>
<td>260</td>
<td>447</td>
</tr>
<tr>
<td>Drive #974</td>
<td>178</td>
<td>674</td>
</tr>
<tr>
<td>Drive #696</td>
<td>33</td>
<td>872</td>
</tr>
<tr>
<td>Drive #969</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>Drive #690</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Drive #680</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Drive #959 contained consumer credit applications.
Second-order analysis uses the *multi-drive correlation*

\[
D = \# \text{ of drives} \nonumber
\]

\[
F = \# \text{ of extracted features} \nonumber
\]

\[
d_0 \ldots d_D = \text{Drives in corpus} \nonumber
\]

\[
f_0 \ldots f_F = \text{Extracted features} \nonumber
\]

\[
FP(f_n, d_n) = \begin{cases} 
0 & f_n \text{ not present on } d_n \\
1 & f_n \text{ present on } d_n
\end{cases} \nonumber
\]

**Scoring Function:**

\[
S_1(d_1, d_2) = \sum_{n=0}^{F} FP(f_n, d_1) \times FP(f_n, d_2) \nonumber
\]
Graph of scoring function:
Graph of scoring function:

The three correlated drives have an extrinsic relationship. (180 drive corpus)
The correlation between Drives #171 and #172 tells a story...

Drive #171: Development drive
- Has source code.
- 346 CCNS; 81 unique.

Drive #172: Production system.
- 31,348 CCNS; 11,609 unique
- Oracle database (hard to reconstruct).

...The programmers used live data to test their system.
Other CCN correlations

#74, #77  Same college in Pacific Northwest. Correlated on CCN “false positive.”

#339 – #356  All used by same New York travel agency

#716, #718  Both from Union City, CA dealer

#814, #820  Both from same Stamford, CT dealer

In two cases, cross-drive correlation discovered drive cataloging errors!
SSN correlation: identical documents on different drives

SSN₁  #342, #343, #356  “Thanks, Laurie” memo

SSN₂  #350, #355  “great grandchildren” memo

But ignore these numbers:

666-66-6666  #313, #427, #429, #430, #612, #627, #744, #770, #808

123-45-6789  #328, #343, #345, #350, #351, #700

555-55-5555  #612, #690
Possible reasons for the same SSN found on two drives

• Two copies of the same document
• Two documents about the same person
• Accidental mismatch

Chance of a false match is 1 in $10^9$. 
Future Work 1: What is the best scoring function?

\[ S_1(d_1, d_2) = \sum_{n=0}^{F} FP(f_n, d_1) \times FP(f_n, d_2) \]
Discount features that appear on many drives

\[ DC(f) = \sum_{n=0}^{D} FP(f, d_n) \]

= \# of drives with feature \( f \)

\[ S_2(d_1, d_2) = \sum_{n=0}^{F} \frac{FP(f_n, d_1) \times FP(f_n, d_2)}{DC(f_n)} \]
Weigh features that are rare on some drives, but high on others

\[ DC(f) = \# \text{ of drives with feature } f \]
\[ FC(f, d) = \text{ count of feature } f \text{ on drive } d \]
\[ S_3(d_1, d_2) = \sum_{n=0}^{F} \frac{FC(f_n, d_1) \times FC(f_n, d_2)}{DC(f_n)} \]
More Future Work:

- Scaling cross-drive correlation to 10,000 drives.
- More sophisticated feature extraction based on Sleuth Kit.
- Use of sector hashes (MD5) to find fragments of documents on different drives.
- Combining CDA with carving and time line analysis.
- Automatically sanitize personal information for publication.
Acknowledgments

• Abhi Shelat (CCN) and Ben Gelb (email)
• Steve Bauer, Gene Spafford, Brian Carrier
• Basis Technology
• University of Auckland
• Harvard University CRCS
Summary

Large-scale forensics is an important problem

Feature Extraction and Cross-drive analysis allow:

- Better single-drive tools
- Intelligent stop-lists
- Identification of social networks

Questions?