Digital Forensics Overview

Garfinkel, Simson L.
Monterey, California. Naval Postgraduate School

http://hdl.handle.net/10945/44323

Downloaded from NPS Archive: Calhoun
Digital Forensics

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School
Feb 7, 2014
http://www.forensicswiki.org/
http://simson.net/
NPS is the Naval Postgraduate School

Monterey, CA — 1500 students
- US Military & Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)

Graduate Schools of Operational & Information Sciences (GSOIS)
- Computer Science
- Defense Analysis
- Information Sciences
- Operations Research
- Cyber Academic Group

National Capital Region (NCR) Office
- 900 N Glebe (Ballston)/Virginia Tech building
- Courses at Ft. Meade

“Evaluation”
• Trusted hardware and software
• Cloud computing

“Exploitation”
• MEDEX — “Media” — Hard drives, camera cards, GPS devices.
• CELEX — Cell phone
• DOCEX — Documents
• DOMEX — Document & Media Exploitation

Current Partners:
• Law Enforcement (FBI & Local)
• DHS (HSARPA; Video Games & Insider Threat
• NSF (Courseware development)
• DOD
Snapchat is a popular app!
Snapchat promised users that expired images could not be viewed unless “saved.”
Snapchat promised users that expired images could not be viewed unless “saved.”

Is there any way to view an image after the time has expired?
No, snaps disappear after the timer runs out. You can save snaps that you capture by pressing the save button on the preview screen.
OMG! — Expired images not actually deleted. They were just hidden from view.

The premise of Snapchat is simple: Send a photo or short video to a friend, and it will self-destruct after 10 seconds. That way, it won’t wind up on the Internet and ruin anyone’s reputation, friendships, or career.

 Needless to say, that has made it a wildly popular choice for sexting. But Snapchat’s appeal goes far beyond that. In an age in which “privacy” and “technology” have become almost antonymous, it has been billed as the anti-Facebook—a communications tool that deletes your data rather than preserving, analyzing, and trading on it. In short, it’s supposed to make messaging fun again.

But the app’s security has never been ironclad. As the media have repeatedly warned parents, and parents in turn warned their kids, message recipients can still save a compromising image by taking a quick screenshot. But Snapchat tries to mitigate the risk somewhat by automatically notifying the sender when that happens. If someone screenshots you, it’s a virtual slap in the face. If they don’t, you can assume you’re in the clear.

Except that apparently you can’t. KSL-TV in Utah reports that an Orem-based firm called Decipher Forensics has figured out a way to recover supposedly deleted images from the recipient’s phone. The process isn’t simple: 24-year-old Decipher forensics examiner Richard Hickman told the network that it takes him about six hours, on average, to image the phone’s data. So far he can only do it with Android devices, though he’s working on doing the same for iOS. But his firm is now offering to perform the recovery procedure for anyone who wants it. From parents
This talk presents today’s digital forensic challenges and presents approaches for solving the triage problem.

Introducing digital forensics

Today’s digital forensics challenges

Approaches for solving the triage problem
Digital information is pervasive in today’s society.

Many potential sources of digital information:
- Desktops; Laptops
- Tablets; Cell Phones
- Internet-Based Services
- Cars

Users of forensic tools have many different goals:
- Document a conspiracy (stock fraud; murder-for-hire)
- Investigation, intelligence & analysis
- Establish possession of contraband information (images; movies)
- Recover “lost” information
- Understand and correct privacy leaks
Many devices record information in non-obvious ways.
Many devices record information in non-obvious ways.
Garmin maps show where you are...

(taken with Garmin’s screen capture)
Map

Map Detail
- Normal

Map View
- North Up

Vehicle
- Change

Trip Log
- Hide

Info
Restore
Maps can also show where you have been.
Maps can also show where you have been.
Acquiring Satellites
Ready to Navigate
Ready to Navigate
Ready to Navigate
The Garmin GPS appears as an external USB storage device with directories and files.
The file 24.gpx contains track information in XML format.

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no" ?><gpx xmlns="http://www.topografix.com/GPX/1/1"
creator="nüvi 1390" version="1.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
```
Accuracy of 0.000001° lat is ≈ 40,000 km ÷ 360 x 0.000001 ≈ 0.1 m
≈ 10cm

GPS accuracy is 7.8 meters w/ 95% confidence level

• http://www.gps.gov/systems/gps/performance/accuracy/
The Garmin GPS stores a lot of information.

Each XML record is roughly 266 bytes:

```
```

The file 24.gpx has 766,553 ÷ 266 ≈ 2,900 tracking points in it.

- My GPS has 44MB of track points from March 7, 2013 ➔ Sept. 15, 2013
  — That’s over 150,000 tracking points!

Inadvertently collected, but incredibly useful.
LandAirSea Magnetic Wireless Pocket-Sized Tracking Key GPS System sold by Amazon: $139.00

Product Features
• Logs the driving speed every second, showing the maximum speed driven for the day.
• Logs the precise GPS location, date and time of every stop
• Software animates the car driving over a digital street map.
• Displays location, date, and time of every stop.
• Records the driving speed, every second. Shows you the maximum speed driven for the day.

See more product details
George Ford’s wife put a LandSeaAir tracker on his car because she suspected he was having an affair. Data collected by the device was instrumental in his murder conviction.
Digital forensics makes *digital evidence* available for decisions

**Preparation:** policy, training & tools

**Collect & preserve evidence**

**Extract preserved data**

**Reporting & Testimony**

**Analysis**
Digital Forensics “research” traditionally focused on training, collection & extraction

Preparation: policy, training & tools
Collect & preserve evidence
Extract preserved data

Training the force
Collection Techniques
Reverse engineering
Most tools were developed for law enforcement.

Tools follow a simple model for manual analysis:

- Extract data
- Make data visibility
- Simple string search

Extraction tools

“Write Blocker”

Hard Drive from desktop

Encase Forensic
Collected data increases dramatically each year.

### FY 12 ACCOMPLISHMENTS

#### BY THE NUMBERS

#### DIGITAL FORENSIC SERVICES

<table>
<thead>
<tr>
<th>Service</th>
<th>FY 12</th>
<th>FY 11</th>
<th>% CHANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agency Requests</td>
<td>842</td>
<td>766</td>
<td>10%</td>
</tr>
<tr>
<td>No. of agencies that received RCFL assistance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service Requests</td>
<td>5,060</td>
<td>6,318</td>
<td>-20%</td>
</tr>
<tr>
<td>No. of requests for assistance received by RCFLs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Examinations Completed</td>
<td>8,566</td>
<td>7,629</td>
<td>12%</td>
</tr>
<tr>
<td>No. of examinations completed at RCFLs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terabytes Processed</td>
<td>5,986</td>
<td>4,263</td>
<td>40%</td>
</tr>
<tr>
<td>A terabyte (TB) is a unit of measurement for data storage capacity equivalent to 1,024 gigabytes. One TB is roughly equivalent to the information in 1,000 encyclopedias.</td>
<td>4,263</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Field Services</td>
<td>553</td>
<td>689</td>
<td>-20%</td>
</tr>
<tr>
<td>No. of onsite operations conducted by law enforcement for which RCFLs provided assistance</td>
<td>689</td>
<td>12%</td>
<td></td>
</tr>
<tr>
<td>Examiners Testimony in Court</td>
<td>101</td>
<td>97</td>
<td>4%</td>
</tr>
<tr>
<td>No. of times RCFL Examiners testified in court and/or at hearings. This does not include subpoenas to testify when testimony was not required</td>
<td>97</td>
<td>11%</td>
<td></td>
</tr>
</tbody>
</table>

#### KIOSK SERVICES

<table>
<thead>
<tr>
<th>Service</th>
<th>FY 12</th>
<th>FY 11</th>
<th>% CHANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Kiosk Use</td>
<td>13,556</td>
<td>8,553</td>
<td>58%</td>
</tr>
<tr>
<td>No. of times law enforcement officers used the CPIK, LMK, and VCPK</td>
<td>8,553</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Phone Investigative Kiosk (CPIK) Use</td>
<td>8,795</td>
<td>5,956</td>
<td>48%</td>
</tr>
<tr>
<td>No. of times law enforcement officers used the CPIK to review cellular phones at RCFLs in support of investigations</td>
<td>5,956</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loose Media Kiosk (LMK) Use</td>
<td>3,665</td>
<td>1,683</td>
<td>118%</td>
</tr>
<tr>
<td>No. of times law enforcement officers used the LMK to review loose media at RCFLs in support of investigations</td>
<td>1,683</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual Cell Phone Kiosk (VCPK) Use</td>
<td>1,096</td>
<td>914</td>
<td>20%</td>
</tr>
<tr>
<td>No. of times law enforcement officers used the VCPK (only available through the HARCFL) to remotely review cellular telephones from their agencies in support of investigations</td>
<td>914</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
My work focuses on developing better approaches for automation and analysis — “big data for little devices”

Automatically identify high-value data.
— Contacts, calendar, documents
— Software
— GPS / Temporal

Correlate — devices with identical or similar copies.
• previously unknown organizations or networks
• data/threats that are unusual or emerging

Presentation and Integration:
• Make the results understandable.
• Effect organizational change through adoption & integration

Human Language Technology
• Apply to العربية, español, 汉语/漢語, 日本語, svenska, etc.
Three principles underly this research.

1. Work with “big data.”
   - “Big data” is our advantage — use it!
   - Many techniques developed on small systems don’t scale
Three principles underly this research.

1. Work with “big data.”
2. Automation is essential.
   • Today most forensic analysis is done manually.
   • We are developing techniques & tools to allow automation.
Three principles underly this research.

1. Work with “big data.”
2. Automation is essential.
3. Concentrate on bulk data.
   - Leverage data that are fragmented and incomplete
     —Deleted and partially overwritten files
     —Fragments of memory in swap & hibernation
     —Tool marks

—Sencar & Memon, ”Identification and recovery of JPEG files with missing fragments,” DFRWS 2009
Example: Integrating Human Language Technology with Digital Forensics

Problems:
- Forensic examiners spend significant time looking for “unusual” file names.
- Many of the file names are not in English.

Solution:
- Model of what makes file names “unusual.”
- Translate non-English file names into English.

New problem: path names are ambiguous and frequently multi-lingual.
- Documents and Settings/defaultuser/ Mes documents/Ma musique/Desktop.ini
- Mis Documentos/SalvadorJP/Excel/ GRUPOS.xls
- Documents and Settings/3742008/ Configuración local/Datos de programa/ Microsoft/Internet Explorer/.
- top.com/تصميماتي/السلسلة المعلوماتية.jpg
  becomes:
  top.com/My designs/The computer-based series.jpg

We do science with “real data.”

The Real Data Corpus (60TB)
• Disks, camera cards, & cell phones purchased on the secondary market.
• Most contain data from previous users.
• Mostly acquire outside the US:
  —Canada, China, England, Germany, France, India, Israel, Japan, Pakistan, Palestine, etc.
• Thousands of devices (HDs, CDs, DVDs, flash, etc.)

Mobile Phone Application Corpus
• Android Applications; Mobile Malware; etc.

The problems we encounter obtaining, curating and exploiting this data mirror those of national organizations
—Garfinkel, Farrell, Roussev and Dinolt, Bringing Science to Digital Forensics with Standardized Forensic Corpora, DFRWS 2009. BEST PAPER AWARD.
—http://digitalcorpora.org/
We manufacture data that can be freely redistributed.

Files from US Government Web Servers (500GB)

- ≈1 million heterogeneous files
  - Documents (Word, Excel, PDF, etc.); Images (JPEG, PNG, etc.)
  - Database Files; HTML files; Log files; XML
- Freely redistributable; Many different file types
- This database was surprising difficulty to collect, curate, and distribute:
  - Scale created data collection and management problems.
  - Copyright, Privacy & Provenance issues.

Advantage over flickr & youtube: persistence & copyright

http://digitalcorpora.org/
Challenges Facing Digital Forensics
Extracting digital evidence was simple five years ago.

“Imaging tools” extracted data without modification.

“Write Blocker” prevents accidental overwriting.

Forensic copy (“disk image”) stored on a storage array.

Original device stored in evidence locker.
Analyzing digital evidence was simple five years ago.

Commercial tools extracted *files* from disk images

- Display of *allocated* & *deleted* files.
- String search
- File extraction
- File “carving”
- Examining disk sectors

Job of analyst:

- Find interesting data
- Report on it.
Today Digital Forensics is faced with 5 crippling challenges:

1. Device Diversity

2. Data Diversity

3. Data Scale

4. Human Capital

5. Cloud & Encryption
Mobile devices exhibit multiple challenges.

Operating system:
- Android? iPhone? Blackberry? Feature Phone?

Access to the data:
- PIN lock?
- Encrypted Storage?
- Stored locally or in the cloud?

Applications:
- Built-in? Downloaded from “App Store”?
- Custom-written?
- Self-destruct / remote wipe?
- Malware?

This $12 phone from Hong Kong could contain important evidence.

<table>
<thead>
<tr>
<th>Spec</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>$12</td>
</tr>
<tr>
<td>CPU</td>
<td>260 MHz, 32-bit</td>
</tr>
<tr>
<td>RAM</td>
<td>8MiB</td>
</tr>
<tr>
<td>Interfaces</td>
<td>USB, microSD, SIM</td>
</tr>
<tr>
<td>Wireless</td>
<td>Quadband GSM, Bluetooth</td>
</tr>
<tr>
<td>Power</td>
<td>Li-Poly battery, includes adapter</td>
</tr>
<tr>
<td>Display</td>
<td>Two-color OLED</td>
</tr>
</tbody>
</table>

http://www.bunniestudios.com/blog/?p=3040
The “CSI Effect” creates unrealistic expectations.

TV digital forensics:
• Every investigator is trained on every tool
• Correlation is easy and instantaneous
• There are no false positives
• Overwritten data can be recovered
• Encrypted data can usually be cracked
• It is impossible to delete anything

The reality:
• Overwritten data cannot be recovered
• Encrypted data usually can't be decrypted
• Forensics rarely answers questions or establishes guilt
• Tools crash a lot

—Digital Forensics: a difficult process that looks easy
Digital Forensics must respond with new science.

Current approaches don’t scale

• User spent *years* assembling email, documents, etc.
• Analysts have days or hours to process it
• Police analyze top-of-the-line systems
  —with *top-of-the-line systems*
• National Labs have large-scale server farms
  —to analyze huge collections

Our approach: leverage our massive data advantage

• Outlier detection and correlation
• Operate autonomously on incomplete, heterogeneous datasets
• Automatically recalibrate; no false positives
The Triage Problem
Digital media triage:  
Deciding where to start

Imagine you encounter a large number of computers, USB drives, etc.

Where do you start?
We have developed many triage techniques.

1. Histogram analysis  
   — *We count the results*
   
   ![Histogram results](image1)

2. Optimistic decompression  
   — *We try to decompress the data*
   
   ![Optimistic decompression example](image2)

3. Cross-drive analysis  
   — *We look for two computers with the same email address*
   
   ![Cross-drive analysis](image3)

4. Sector hashing  
   — *We look at individual disk sectors.*
   
   ![Sector hashing](image4)

5. Random sampling  
   — *We look at 1% of the data, randomly chosen*
   
   ![Random sampling](image5)
Email addresses can reveal:

- User(s) of a device
- Associates
- Connections between devices

Today’s forensic tools implement two strategies for extracting email addresses.

1. Text extraction from files
2. Text extraction from bulk data
But most digital forensics tools find email addresses with string search.

With string search, you only find what you are looking for.
bulk_extractor extracts email addresses.

For each email address, reports location, email address, & context

34427974 grafta@bl.com 
24900678 grafta@bl.com 
50392739 inet@microsoft.com 
26735686 grafta@bl.com 
39265456 domexuser2@gmail.com 
39267100 domexuser2@live.com 
39269992 domexuser1@gmail.com 
39270105 domexuser1@gmail.com 
40893040 domexuser2@live.com 
40948912 domexuser2@gmail.com 
40950441 domexuser2@live.com 
51781228 dbaron@dbaron.org 
51788157 bzbarsky@mit.edu 
51789901 bzbarsky@mit.edu
Histogram analysis shows the important email addresses.

- domexuser1@gmail.com (utf16=303)
- domexuser2@gmail.com (utf16=225)
- domexuser3@gmail.com (utf16=204)
- ips@mail.ips.es
- premium-server@thawte.com
- cps-requests@verisign.com (utf16=3)
- someone@example.com (utf16=234)
- domexuser2@live.com (utf16=59)
- domexuser1@hotmail.com (utf16=80)
- domexuser1@live.com (utf16=59)
- domexuser2@hotmail.com (utf16=97)
- inet@microsoft.com
- example@passport.com (utf16=115)
- myname@msn.com (utf16=115)
- info@valicert.com
- piracy@microsoft.com (utf16=91)
- certificate@trustcenter.de
- name_123@hotmail.com (utf16=78)
- talkback@mozilla.org (utf16=12)
- hewitt@netscape.com (utf16=1)
- lord@netscape.com

Notice that we combine UTF8 and UTF16
It’s important to distinguish email addresses that are relevant to a case from those that are not.

The #4 address is ips@mail.ips.es

- We should probably ignore these

Other sources that to ignore:

- Windows binaries; SSL certificates; Sample documents; News Stories

Ignore lists are expensive to maintain.

- Find them automatically by analyzing multiple drives!
- Email addresses on thousands of different drives are probably irrelevant.

<table>
<thead>
<tr>
<th>Address</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:domexuser1@gmail.com">domexuser1@gmail.com</a></td>
<td>609</td>
</tr>
<tr>
<td><a href="mailto:domexuser2@gmail.com">domexuser2@gmail.com</a></td>
<td>455</td>
</tr>
<tr>
<td><a href="mailto:domexuser3@gmail.com">domexuser3@gmail.com</a></td>
<td>359</td>
</tr>
<tr>
<td><a href="mailto:ips@mail.ips.es">ips@mail.ips.es</a></td>
<td>268</td>
</tr>
<tr>
<td><a href="mailto:premium-server@thawte.com">premium-server@thawte.com</a></td>
<td>252</td>
</tr>
<tr>
<td><a href="mailto:cps-requests@verisign.com">cps-requests@verisign.com</a></td>
<td>243</td>
</tr>
<tr>
<td><a href="mailto:someone@example.com">someone@example.com</a></td>
<td>243</td>
</tr>
<tr>
<td><a href="mailto:domexuser2@live.com">domexuser2@live.com</a></td>
<td>221</td>
</tr>
<tr>
<td><a href="mailto:domexuser1@hotmail.com">domexuser1@hotmail.com</a></td>
<td>198</td>
</tr>
<tr>
<td><a href="mailto:domexuser1@live.com">domexuser1@live.com</a></td>
<td>185</td>
</tr>
</tbody>
</table>
Storage devices arrange files in individual sectors.

Email addresses might be in any file, any sector:
Every email address is a sequence of bytes.

A simple email address:

\[ \text{XYZ@company.com} \]

Stored on disk / in memory as 15 bytes:

\[ \text{XYZ@company.com} \]

Each byte is 8-bits. Range is 0-255

\[ \begin{array}{ccccccccccccccc}
\end{array} \]

Normally bytes are displayed in hexadecimal notation:

\[ \begin{array}{ccccccccccccccc}
58 & 59 & 5a & 40 & 63 & 6f & 6d & 70 & 61 & 6e & 79 & 2e & 63 & 6f & 6d \\
\end{array} \]

This is called UNICODE (UTF-8)
Every email address is a sequence of bytes.

A simple email address:

XYZ@company.com

Stored on disk / in memory as 15 bytes:

XYZ@company.com

Each byte is 8-bits. Range is 0-255

88 89 90 64 99 111 109 112 97 110 121 46 99 111 109

Normally bytes are displayed in hexadecimal notation:

58 59 5a 40 63 6f 6d 70 61 6e 79 2e 63 6f 6d

This is called UNICODE (UTF-8)
Some email addresses are easy to spot

| a097 83a1 ed96 26a6 3c69 3d0f 750a 2399 | ......&.<i=..u.#. |
| a2b5 bea7 692f 5847 a38a dd53 082c add5 | ...i/XG...S..., |
| 5061 b64c 721d 864b 90b6 b55f bb04 735c | Pa.Lr..K...__...s\ |
| 9448 6730 5453 df64 813e b603 5795 2242 | .Hg0TS.d..>..W."B |
| e9c8 7454 7322 7cdc b60e 97af 2f64 2728 | ..tTs"|....../d' |
Email addresses like these can be found with regular expressions

<table>
<thead>
<tr>
<th>a097 83a1 ed96 26a6 3c69 3d0f 750a 2399</th>
<th>a2b5 bea7 692f 5847 a38a dd53 082c add5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5061 b64c 721d 864b 90b6 b55f bb04 135c</td>
<td>Pa.Lr..K..._...s\</td>
</tr>
<tr>
<td>9448 6730 5453 df64 813e b603 5795 1142</td>
<td>.Hg0TS.d&gt;..W.&quot;B</td>
</tr>
</tbody>
</table>
| e9c8 7454 7322 7cdc b040 7774 723a 723a | ..tTs"| ....../d'
| 3c58 595a 4043 4f4d 5f4d 7f4d 7f4d 7f4d | <XYZ@COMPANY.COM |
| a9e9 e92c a3f8 6e46 0946 4946 4946 4946 | ...,...nF.0.....]+ |
| d89d 77cc fele f637 f313 dvar 1b47 05eb | ..w....7......G.. |

**XYZ@company.com**
Problem: byte sequences can be encoded in many ways.

**XYZ@company.com**

- **Unicode:** “XYZ@company.com”
  
  58 59 5a 40 63 6f 6d 70 61 6e 79 2e 63 6f 6d

- **Base 16:** “58595a40636f6d70616e792e636f6d0a”
  
  3538 3539 3561 3430 3633 3666 3664 3730 3631 3665 3739 3265 3633 3664 3061 616e792e636f6d0a

- **Base 64:** “WFlaQGNvbXBhbnkuY29tCg==”
  
  5746 6c61 5147 4e76 6258 4268 626e 6b75 5932 3974 4367 3d3d 3d0a

- **Composition:** `echo “XYZ@company.com” | compress | xxd`
  
  1f9d 9058 b268 0132 e64d 1b38 61dc e471 ...X.h.2.M.8a..q
  
  51b0 8d02 Q...
Compression works by eliminating repeated sequences:

Computers use compression to save memory:

5859 5a40 636f 6d70 616e 792e 636f 6d  
ABC@company.com
4142 4340 636f 6d70 616e 792e 636f 6d  
DEF@company.com
4445 4640 636f 6d70 616e 792e 636f 6d  
XYZ@company.com

Compressed with “gzip:”

1f8b 0800 0000 0000 0203 8b88 8c72 48ce  
.............rH.
cf2d 48cc abd4 03d2 0a8e 4ece 287c 1757  
.-H.......N.(|.W
3714 3e00 b455 c1c5 3000 0000  
7..U..0...

Compressed email addresses do not “look” like email addresses!

—You can’t find them with regular expressions!
It’s hard to see compressed email address in bulk data.
It’s hard to see compressed email address in bulk data.

- XYZ@company.com
- ABC@company.com
- DEF@company.com

Folders.pst
Presentation.pptx
Sequestration.docx
Mother.JPG
### bulk_extractor breaks the disk into individual blocks.

<table>
<thead>
<tr>
<th>Block Hash</th>
<th>File Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>e327 962d 6450 3d91 1f8b 0800 0000 0000 cf2d 48cc abd4 03d2 3714 3e00 b455 c1c5 0a8e 4ece 287c 1757</td>
<td>Folders.pst</td>
</tr>
<tr>
<td>c945 3bed 97a6 a4cd 0203 8b88 8c72 48ce 0a8e 4ece 287c 1757 3000 0000 0000 0000 3714 3e00 a175 10ed</td>
<td>Sequestration.docx</td>
</tr>
<tr>
<td>3714 3e00 b455 c1c5 0a8e 4ece 287c 1757</td>
<td>Presentation.pptx</td>
</tr>
<tr>
<td>a097 83a1 ed96 26a6 a2b5 bea7 692f 5847 5061 b64c 721d 864b 9448 6730 5453 df64 e9c8 7454 7322 7cdc 3cfb 84bd 2a84 2dfe a9e9 e92c a3f8 6e46 d89d 77cc fe1e f637</td>
<td>Sequestration.docx</td>
</tr>
<tr>
<td>c945 3bed 97a6 a4cd 0203 8b88 8c72 48ce 0a8e 4ece 287c 1757 3000 0000 0000 0000 3714 3e00 a175 10ed</td>
<td>Sequestration.docx</td>
</tr>
</tbody>
</table>

- Folders.pst
- Presentation.pptx
- Sequestration.docx
- Mother.JPG

---

<XYZ@COMPANY.COM>
Each block is tested for compressed data. Within each block, we check each byte.

<table>
<thead>
<tr>
<th>TEST 1</th>
<th>TEST 2</th>
<th>TEST 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>e327 962d 6450 3d91</td>
<td>c945 3bed 97a6 a4cd</td>
<td>.'-dP=.E; ......</td>
</tr>
<tr>
<td>1000 0000 0000 0000</td>
<td>0a8e 4ece 217c 2157</td>
<td>-H(1.</td>
</tr>
<tr>
<td>3714 3e00 b455 c1c5</td>
<td>3714 3e00 a175 10ed</td>
<td>..N.(.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST 4</th>
<th>TEST 5</th>
<th>TEST 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a097 83a1 ed96 26a6</td>
<td>3c69 3d0f 750a 2399</td>
<td>......&amp;.&lt;i=.#.</td>
</tr>
<tr>
<td>a2b5 bea7 682f 5847</td>
<td>a38a dd53 082c add5</td>
<td>...i/XG...S....</td>
</tr>
<tr>
<td>5041 5047 72d 874b</td>
<td>90b5 9754 98b4 735c</td>
<td>Fh..__._..</td>
</tr>
<tr>
<td>9448 6730 5435 df64</td>
<td>813e b003 3795 2242</td>
<td>_Hg0V3.a:.&gt;..W..B</td>
</tr>
<tr>
<td>e9c8 7454 7322 7cdc</td>
<td>b60e 97af 2f64 2728</td>
<td>..tTs&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST 7</th>
<th>TEST 8</th>
<th>TEST 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>a9e9 b356 a3f8 5a6</td>
<td>50ea 5935 c349 1513</td>
<td>&lt;<a href="mailto:XYZ@COMPANY.COM">XYZ@COMPANY.COM</a></td>
</tr>
<tr>
<td>39b6 2a84 2dfe</td>
<td>0500 c88c 37a2 5599</td>
<td>...[0..]</td>
</tr>
<tr>
<td>d89d 77cc fele 1637</td>
<td>f313 dval 1b47 c09b</td>
<td>...w.../......G..</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST A</th>
<th>TEST B</th>
<th>TEST C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folders.pst</td>
<td>Presentation.pptx</td>
<td>Sequestration.docx</td>
</tr>
<tr>
<td>Mother.JPG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
bulk_extractor run on a disk from India (IN10-0138) found many compressed email addresses.

<table>
<thead>
<tr>
<th>Encoding</th>
<th>count</th>
<th>1) Plain in Files</th>
<th>2) Comp. in Files</th>
<th>3) Plain in non-</th>
<th>4) Comp in non-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleartext</td>
<td>358</td>
<td>--</td>
<td>--</td>
<td>5341</td>
<td>--</td>
</tr>
<tr>
<td>All Comp</td>
<td>--</td>
<td>9</td>
<td>--</td>
<td>--</td>
<td>135</td>
</tr>
<tr>
<td>GZIP</td>
<td>50</td>
<td>13</td>
<td>1</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>HIBER</td>
<td>39</td>
<td>6</td>
<td>1</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>HIBER-GZIP</td>
<td>23</td>
<td></td>
<td></td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>PDF</td>
<td>88</td>
<td>1</td>
<td></td>
<td>9</td>
<td>78</td>
</tr>
<tr>
<td>ZIP</td>
<td>28</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>ZIP-PDF</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>18</td>
</tr>
</tbody>
</table>

135 out of 5700 email addresses are invisible to other forensic tools.
We extended the analysis to 1,646 disk images and many codings.

<table>
<thead>
<tr>
<th>Coding</th>
<th>Drives</th>
<th>Emails</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Plain in files</td>
<td>739</td>
<td>81,920</td>
<td>4,206</td>
</tr>
<tr>
<td>2) Comp in files</td>
<td>355</td>
<td>19,711</td>
<td>5,454</td>
</tr>
<tr>
<td>3) Plain in non-files</td>
<td>860</td>
<td>1,956,059</td>
<td>178,073</td>
</tr>
<tr>
<td>4) Comp in non-files</td>
<td>474</td>
<td>165,481</td>
<td>59,376</td>
</tr>
<tr>
<td>BASE64 Comp</td>
<td>54</td>
<td>219</td>
<td>50</td>
</tr>
<tr>
<td>BASE64-GZIP Comp</td>
<td>2</td>
<td>64</td>
<td>37</td>
</tr>
<tr>
<td>GZIP Comp</td>
<td>234</td>
<td>66,195</td>
<td>9,103</td>
</tr>
<tr>
<td>GZIP-BASE64 Comp</td>
<td>7</td>
<td>44</td>
<td>11</td>
</tr>
<tr>
<td>GZIP-GZIP Comp</td>
<td>15</td>
<td>12,663</td>
<td>11,845</td>
</tr>
<tr>
<td>GZIP-GZIP-BASE64 Comp</td>
<td>2</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>GZIP-GZIP-GZIP Comp</td>
<td>4</td>
<td>58</td>
<td>38</td>
</tr>
<tr>
<td>GZIP-GZIP-ZIP Comp</td>
<td>1</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>GZIP-PDF Comp</td>
<td>5</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>GZIP-ZIP Comp</td>
<td>6</td>
<td>49</td>
<td>30</td>
</tr>
<tr>
<td>HIBER Comp</td>
<td>79</td>
<td>1,433</td>
<td>217</td>
</tr>
<tr>
<td>PDF Comp</td>
<td>162</td>
<td>2,352</td>
<td>238</td>
</tr>
<tr>
<td>ZIP Comp</td>
<td>388</td>
<td>85,252</td>
<td>59,369</td>
</tr>
<tr>
<td>ZIP-BASE64 Comp</td>
<td>5</td>
<td>30</td>
<td>13</td>
</tr>
<tr>
<td>ZIP-BASE64-GZIP Comp</td>
<td>2</td>
<td>65</td>
<td>38</td>
</tr>
<tr>
<td>ZIP-GZIP Comp</td>
<td>14</td>
<td>261</td>
<td>132</td>
</tr>
<tr>
<td>ZIP-PDF Comp</td>
<td>26</td>
<td>115</td>
<td>18</td>
</tr>
</tbody>
</table>
We can scan for credit card numbers (CCNs), phone numbers, addresses, and other structured info.

We use multiple filters to minimize false positives

**Disk #105:**

<table>
<thead>
<tr>
<th>Test</th>
<th># pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>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>patterns &amp; histogram</td>
<td>38</td>
</tr>
</tbody>
</table>

Sample output:

| 13152133 | 'CHASE NA | 5422-4128-3008-3685 |
| 13152440 | 'DISCOVER | 6011-0052-8056-4504 |
| 13152589 | 'GE CARD  | 4055-9000-0378-1959 |
| 13152740 | BANK ONE   | 4332-2213-0038-0832 |
| 13153182 | 'NORWEST   | 4829-0000-4102-9233 |
| 13153332 | 'SNB CARD  | 5419-7213-0101-3624 |
Most drives had just a few, but some had a lot of credit card numbers.
Most drives had just a few, but some had a lot of credit card numbers.

What would it mean if two drives had a lot of credit card numbers in common?
Scenario #1: The owner of one drive sent a message to another drive.
Scenario #2:
Both drives received a message from a third party.
Cross Drive Analysis (CDA) computes the correlation matrix of the distinguishing information.

<table>
<thead>
<tr>
<th></th>
<th>first drive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>second drive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

High correlation
This correlation of 250 drives automatically identified media from the same organization.

This correlation of 250 drives automatically identified media from the same organization.

Sector hashing uses fragments of JPEGs as identifiers.

A relatively small JPEG:

Bytes: 31,046
JPEG files have internal structure

Bytes: 31,046

Header
Icons
EXIF
Color Table
Huffman Encoded Data
Footer

[FF D8 FF E0] or [FF D8 FF E1]

[FF D9]
This JPEG has 61 sectors.

---

Garfinkel, Simson, Vassil Roussev, Alex Nelson and Douglas White, *Using purpose-built functions and block hashes to enable small block and sub-file forensics*, DFRWS 2010, Portland, OR
This JPEG has 61 sectors. 41 of the sectors are “distinct” — not repeated elsewhere.

Each BLUE sector is distinct.

- Each has a distinct sequence of bytes
  
dd6a d66f baac 8660 829b b59c 329c 006c 4dbc 5884 5104 3d9b fcae d375 1fab 3ff!
  766e 81c0 12f6 b1a0 5bec ab9a f425 943; 02ec bace 23d6 eba0 762b 4b9f 53d0 61de e003 059c f75c dc9c fdd5 63e2 2696 74a;
  ....

- Each has a distinct MD5 “finger print.”
  2b00042f7481c7b056c4b410d28f33cf
If the image is on a hard drive, each “block” of the image occupies one sector.
The sectors are distinct!
∴ A single sector ➔ the image was probably on the drive.

Two modes of operation:
1. Scan for fragments of deleted files.
2. Use random sampling to scan a 1TB drive in 10 minutes
   \( p = .99 \) of finding a sector

Law enforcement is a typical application for this technology.

US agents encounter a hard drive at a border crossings...
- Media can be rapidly searched in a way that respects privacy.
  — 5-10 minutes
- Media can be exhaustively searched.
  — 2-3 hours

Searches turn up rooms filled with servers....
- Systems can be automatically analyzed
- Servers likely to contain evidence can be manually reviewed.
- Connections between servers can be inferred automatically.

Big challenge: tool development and deployment.
Where do we go from here?
For further reading...

**Cover Feature**

**SMARTPHONE SECURITY CHALLENGES**

Jan Wong, Kaveh Foroush, and David Smith, Bahcesehir University

Because of their unique characteristics, smartphones provide challenges for malware analysts that offer countermeasures to help mitigate their security.

**SMARTPHONE THREATS AND ATTACKS**

Many security challenges remain with smartphones because of their unique characteristics. They include the ease of use and the high cost, which make them attractive targets for malware analysts. There are also many security challenges that come from the fact that smartphones are used in the workplace.

**SMARTPHONE SECURITY THREATS**

There are many threats that can undermine smartphone security, including eavesdropping, social engineering, and malware. A recent study by the University of California, Berkeley, found that 40 percent of Android apps had at least one vulnerability.

**SCADA SYSTEMS: CHALLENGES FOR FORENSIC INVESTIGATORS**

Vlad R. Vang, University of New Hampshire

**Sector Hashes for Target File Detection**

Irfan Ahmed, Kevin Streff, and Sonell Raman, Naval Postgraduate School

Sector hashing, digital forensic investigators can use sector-based file detection to determine whether a file is present on a disk. This technique can be used to detect files that have been deleted or that are hidden from traditional file system searches.

**SECTOR HASHING**

Sector hashing is a method for identifying sectors on a file system. It is based on the fact that files are stored on sectors of a disk drive, and that each sector contains a fixed amount of data. To identify a file, the investigator can compare the contents of the file with the contents of a corresponding sector on the disk.

**SCADA SYSTEMS**

SCADA systems are used to control and monitor industrial processes. They include a variety of systems such as power plants, water treatment plants, and chemical plants. SCADA systems are used to control and monitor these processes, and they are connected to physical plants and to electronic systems. They are connected to the Internet through a variety of means, such as Ethernet, Wi-Fi, and cellular networks.

**SECTOR HASHING**

Sector hashing is a method for identifying sectors on a file system. It is based on the fact that files are stored on sectors of a disk drive, and that each sector contains a fixed amount of data. To identify a file, the investigator can compare the contents of the file with the contents of a corresponding sector on the disk.

**SECURITY THREATS**

There are many security threats that can undermine smartphone security, including eavesdropping, social engineering, and malware. A recent study by the University of California, Berkeley, found that 40 percent of Android apps had at least one vulnerability.

**SMARTPHONE SECURITY CHALLENGES**

Jan Wong, Kaveh Foroush, and David Smith, Bahcesehir University

Because of their unique characteristics, smartphones provide challenges for malware analysts that offer countermeasures to help mitigate their security.

**SMARTPHONE THREATS AND ATTACKS**

Many security challenges remain with smartphones because of their unique characteristics. They include the ease of use and the high cost, which make them attractive targets for malware analysts. There are also many security challenges that come from the fact that smartphones are used in the workplace.

**SMARTPHONE SECURITY THREATS**

There are many threats that can undermine smartphone security, including eavesdropping, social engineering, and malware. A recent study by the University of California, Berkeley, found that 40 percent of Android apps had at least one vulnerability.

**SCADA SYSTEMS: CHALLENGES FOR FORENSIC INVESTIGATORS**

Vlad R. Vang, University of New Hampshire

**Sector Hashes for Target File Detection**

Irfan Ahmed, Kevin Streff, and Sonell Raman, Naval Postgraduate School

Sector hashing, digital forensic investigators can use sector-based file detection to determine whether a file is present on a disk. This technique can be used to detect files that have been deleted or that are hidden from traditional file system searches.

**SECTOR HASHING**

Sector hashing is a method for identifying sectors on a file system. It is based on the fact that files are stored on sectors of a disk drive, and that each sector contains a fixed amount of data. To identify a file, the investigator can compare the contents of the file with the contents of a corresponding sector on the disk.

**SCADA SYSTEMS**

SCADA systems are used to control and monitor industrial processes. They include a variety of systems such as power plants, water treatment plants, and chemical plants. SCADA systems are used to control and monitor these processes, and they are connected to physical plants and to electronic systems. They are connected to the Internet through a variety of means, such as Ethernet, Wi-Fi, and cellular networks.

**SECTOR HASHING**

Sector hashing is a method for identifying sectors on a file system. It is based on the fact that files are stored on sectors of a disk drive, and that each sector contains a fixed amount of data. To identify a file, the investigator can compare the contents of the file with the contents of a corresponding sector on the disk.

**SECURITY THREATS**

There are many security threats that can undermine smartphone security, including eavesdropping, social engineering, and malware. A recent study by the University of California, Berkeley, found that 40 percent of Android apps had at least one vulnerability.
Digital Forensics

Finding and preserving evidence of crime in electronic memory requires careful methods as well as technical skill

Simson L. Garfinkel

Once the 1980s, computers have had an increasing role in all aspects of human life—including an involvement in criminal acts. This development has led to the rise of digital forensics, the uncovering and in-depth examination of evidence located on computers or other electronic devices and networks. Because of both the scale and the diversity of their domain, digital forensics researchers and practitioners stand at the forefront of some of the most challenging problems in computer science today, including "big data," analysis of language processing, data visualizations and cybersecurity.

Compared with traditional forensic science, digital forensics poses significant challenges. Information on a computer system can be changed without a trace, the scale of data that must be analyzed is vast, and the variety of data types is enormous. Just as a traditional forensic investigator must be prepared to analyze any kind of smear or fragment, no matter the source, a digital investigator must be able to make sense of any data that might be found on any device anywhere on the planet—a very difficult proposition.

From its inception, digital forensics has served two different purposes, each with its own difficulties. First, in many cases computers contain evidence of a crime that is part of the physical world. The computer was all but incidental—except that computerization has made the evidence harder to investigate. For example, financial examiners often keep a list of things their victims’ accounts using an IBM AS/400 mainframe computer from the 1980s. The age of the computer helped perpetuate the crimes, because two people on Wall Street have experience with 25-year-old technology, and it created an added complication after Madoff was arrested, because in the form of log files and archives, or inadvertently, as a result of software that does not clearly erase memory and files. As a result, investigators can frequently recover old email messages, chat logs, Google search terms, and other kinds of data that were created weeks, months, or even years before the crimes were committed.

The second class of digital forensics is completely electronic. Information on a computer system can be changed without a trace, the scale of data that must be analyzed is vast, and the variety of data types is enormous.

Information on a computer system can be changed without a trace, the scale of data to be analyzed is vast, and the variety of data types is enormous.

Sun Microsystems Laboratories, a division of Oracle Corporation.

Information on a computer system can be changed without a trace, the scale of data to be analyzed is vast, and the variety of data types is enormous.

Sun Microsystems Laboratories, a division of Oracle Corporation.

For further reading...

There’s still a lot of research to do!

Summer 2013: Analysis of XOR obfuscation in the wild
• 4 interns: 1 Poolesville MD High School student (first author) & 3 West Point cadets

Ecological Studies:
— Better understanding of what happens inside SQLite3 database files.
— Improved exploitation of RFC822/2822 “headers” in Email, Web Servers, etc.

Identity analytics and disambiguation
• Identify shared accounts or when an email address passes between users.
• Identification of paired (work, home) accounts.

“Big data” and data mining
• Cross-country synchronization of multiple 1PB data sets.
  — Addressing undetectable read/write errors and silent corruption.
• Identify hostile insiders with outlier analysis

Visualization and Data Fusion:
• Present complex results in simple, straightforward reports.
• Combine stored data, network data, and Internet-based information.

Contact Information:
Simson L. Garfinkel
slgarfin@nps.edu
http://simson.net/
http://digitalcorpora.org/