Zero-Click Security

Simson L. Garfinkel
Center for Research on Computation and Society
Harvard University
March 6, 2006
The Tandy 200
Purchased used from a computer store in August 1998:
HCI-SEC: The merging of security and usability

2003: CRA “Grand Challenge”

2004: IEEE S&P Special Issue

2005: PITAC “priority”

2005: Cranor & Garfinkel Book
Aligning Security and Usability:

Zero-Click, not Zero-Visibility

Frequently requires rethinking and redesigning.
Hidden information is a widespread Usability/Security problem today.

Tandy 200

Hard Drive

USB drive

PDF file
There are roughly a dozen documented cases of people purchasing old PCs and finding sensitive data.

- A woman in Pahrump, NV bought a used PC with pharmacy records [Markoff 97]

- Pennsylvania sold PCs with “thousands of files” on state employees [Villano 02]

- Paul McCartney’s bank records sold by his bank [Leyden 04]

- O&O Software GmbH – 100 drives.[O&O 04]

- O&O Software GmbH – 200 drives.[O&O 05]

None of these are scientifically rigorous studies.
This is a huge problem:
210 million drives were retired in 2005!
There is a significant market for used disk drives.

Retired drives are:

• Re-used within organizations
• Given to charities
• Sold at auction

About 1000 used drives/day sold on eBay.
In 1998 I decided to start purchasing hard drives on the secondary market.

2001: 100 drives
2003: 150 drives
2005: 500 drives
2006: 950 drives
[Garfinkel & Shelat 03] established the scale of the problem.

With 150 hard drives purchased on eBay we found:

- Thousands of credit card numbers
- Financial records
- Medical information
- Trade secrets
- Highly personal information

We did not determine why the data had been left behind.
There are three primary techniques for assuring data confidentiality.

1. Physical security.
2. Logical access controls. (operating system)
3. Cryptography (disk & link)
These techniques don’t work when a disk is thrown out or repurposed.

1. Physical security

2. Logical access controls (operating system)

3. Cryptography (disk & link)

4. (Physical destruction)

Most people don’t encrypt their data.
FORMAT C: doesn’t erase the hard drive.

FORMAT just writes a new root directory.
DEL doesn’t delete files

DEL simply removes the file’s name from the directory.
Drives arrive by UPS and USPS
Drives are “imaged” with **aimage**.
Images stored on external firewire drives

900GB of storage holds 800 hard drive images
Example: Disk #70: IBM-DALA-3540/81B70E32

Purchased for $5 from a Mass retail store on eBay
Copied the data off: 541MB

Initial analysis:

- Total disk sectors: 1,057,392
- Total non-zero sectors: 989,514
- Total files: 3

The files:

- drwxrwxrwx 0 root 0 Dec 31 1979 ./
- r-xr-xr-x 0 root 222390 May 11 1998 IO.SYS
- r-xr-xr-x 0 root 9 May 11 1998 MSDOS.SYS
- rwxrwxrwx 0 root 93880 May 11 1998 COMMAND.COM
Clearly, this disk had been FORMATed...

Windows FORMAT doesn’t erase the disk... FORMAT just writes a new root directory.
UNIX “strings” reveals the disk’s previous contents...

Insert diskette for drive
   and press any key when ready
Your program caused a divide overflow error.
If the problem persists, contact your program vendor.
Windows has disabled direct disk access to protect your long filenames.
To override this protection, see the LOCK /? command for more information.
The system has been halted. Press Ctrl+Alt+Del to restart your computer.
You started your computer with a version of MS-DOS incompatible with this version of Windows. Insert a Startup diskette matching this:

OEMString = "NCR 14 inch Analog Color Display Enchanced SVGA, NCR Corporation"
Graphics Mode: 640 x 480 at 72Hz vertical refresh.
XResolution = 640
YResolution = 480
VerticalRefresh = 72
IBM AntiVirus Trial Edition is a full-function but time-limited evaluation version of the IBM AntiVirus Desktop Edition product. You may have received the Trial Edition on a promotional CD-ROM or as a single-file installation program over a network. The Trial Edition is available in seven national languages, and each language is provided on a separate CC-ROM or as a separate:

EAS.STC
EET.STC
ELR.STC
ELS.STC
MAB-DEDUCTIBLE
MAB-MOOP
MAB-MOOP-DED
METHIMAZOLE
INSULIN (HUMAN)
COUMARIN ANTICOAGULANTS
CARBAMATE DERIVATIVES
AMANTADINE
MANNITOL
MAPROΤILINE
CARBAMAZEPINE
CHLORPHENESIN CARBAMATE
ETHINAMATE
FORMALDEHYDE
MAFENIDE ACETATE
Data left behind in computer systems is a serious social problem.

Large numbers of drives are being sold and given away.

Many of them appear to have hidden confidential information.

Computer Science is morally obligated to solve this problem!
To be effective, a solution must address the root cause

**Usability Problem:**

- Effective audit of information present on drives.
- Make DEL and FORMAT actually remove data. [Bauer & Priyantha 01]
- Provide alternative strategies for data recovery.

**Education Problem:**

- Add training to the interface. [Whitten 04]
- Regulatory requirements. [FTC 05, SEC 05]
- Legal liability.

To find that cause, I looked *on the drives* and *contacted the data subjects.*
Data on a hard drive is arranged in sectors.

The white sectors indicate directories and files that are visible to the user.
Data on a hard drive is arranged in sectors.

The brown sectors indicate files that were deleted.
Data on a hard drive is arranged in sectors.

The green sectors indicate sectors that were never used (or that were wiped clean).
Stack the disk sectors:

```
usr
bin
ls
cp
tmp
slg
/
b
mail
junk
beth
x5
x4
x3
x2
x1
x6
x7
```

Files

Deleted Files

Zero Blocks

Files
NO DATA: The disk is factory fresh.

All Blocks are Zero
The disk has an empty file system
AFTER OS INSTALL: Temp. files have been deleted

Free Blocks

Deleted temporary files

OS and Applications
AFTER A YEAR OF SERVICE

Blocks never written

Deleted files

... 1 year ...

OS, Applications, and user files

Files
Deleted Files
Zero Blocks
time
All Blocks are
Zero
Blank
Blocks
File System Structures

deleted temporary files

... 1 year ...

OS, Applications, and user files
DISK NEARLY FULL!

... 1 year ...

OS, Apps, user files, and lots of MP3s!
FORMAT C:\ (to sell the computer.)

... 1 year ...

Recoverable Data

File System Structures
- Free Blocks
- OS and Applications
- Deleted temporary files
- Deleted files
- Blocks never written
- OS, Apps, user files, and lots of MP3s!

Files
- Deleted Files
- Zero Blocks

Recoverable Data
We can use forensics to reconstruct motivations:

![Diagram showing time, OS, Apps, user files, and lots of MP3s! Recoverable Data, Training failure, and Usability failure.](image)
Drives I collected 1998-2003 are dominated by failed sanitization attempts...

..but training failures are also important.
But what *really* happened?

I needed to contact the original drive owners.
The *Remembrance of Data Passed* Traceback Study.  
[Garfinkel 05]

1. Find data on hard drive

2. Determine the owner

3. Get contact information for organization

4. Find the right person *inside* the organization

5. Set up interviews

6. Follow guidelines for human subjects work

This was a lot harder than I thought it would be.
Ultimately, I contacted 20 organizations between April 2003 and April 2005.
The leading cause: betrayed trust.

Trust Failure: 5 cases

- Home computer; woman’s son took to “PC Recycle”
- Community college; no procedures in place
- Church in South Dakota; administrator “kind of crazy”
- Auto dealership; consultant sold drives he “upgraded”
- Home computer, financial records; same consultant

This specific failure wasn’t considered in [GS 03]; it was the most common failure.
Second leading cause: Poor training and supervision

Trust Failure: 5 cases

Lack of Training: 3 cases

- California electronic manufacturer
- Supermarket credit-card processing terminal
- ATM machine from a Chicago bank

Alignment between the interface and the underlying representation would overcome this problem.
Sometimes the data custodians just don’t care.

Trust Failure: 5 cases
Lack of Training: 3 cases
Lack of Concern: 2 cases

- Bankrupt Internet software developer
- Layoffs at a computer magazine

Regulation on resellers might have prevented these cases.
In seven cases, no cause could be determined.

Trust Failure: 5 cases  
Lack of Training: 3 cases  
Lack of Concern: 2 cases

Unknown Reason: 7 cases

- Bankrupt biotech startup
- Another major electronics manufacturer
- Primary school principal’s office
- Mail order pharmacy
- Major telecommunications provider
- Minnesota food company
- State Corporation Commission

Regulation might have helped here, too.
I have identified five distinct patterns for addressing the sanitization problem.

Naming these patterns is the first step to deployment.
The power of these patterns is that they apply equally well to other sanitization problems.

- Document Files

- Web Browsers
Information is left in document files.

• The New York Times published a PDF file containing the names of Iranians who helped with the 1953 coup. [Young 00]

• US DoJ published a PDF file “diversity report” containing embarrassing redacted information. [Poulsen 03]

• SCO gave a Microsoft Word file to journalists that revealed its Linux legal strategy. [Shankland 04]

• Multinational forces in Iraq published classified information about insurgency methods.
Acrobat is literally a threat to national security.

NSA recently published a “how to sanitize” guide.
Microsoft has tried to solve this problem with its “Remove Hidden Data” tool.
Microsoft has tried to solve this problem with its “Remove Hidden Data” tool.
Microsoft has tried to solve this problem with its “Remove Hidden Data” tool.

My patterns predict that Microsoft’s tool will fail.
The information leaks because two patterns were not implemented.

Current agenda: getting vendors to implement these patterns.
The techniques developed for [Garfinkel ’05] are different than traditional forensics techniques.

Traditional forensics tools:

- Interactive user interface.
- Recovery of “deleted” files.
- Generation of “investigative reports” for courtroom use.
- Focus on one or a few disks.

In [Garfinkel ’05], there were *hundreds* of disks to analyze.
Today’s tools choke when confronted with thousands of disks.

• Has this drive been previously imaged?
• Which drives belong to my target?
• Do any drives belong to my target’s associates?
• Where should I start?

Today’s tools are for criminal investigations. Increasingly, we need tools for intelligence analysis.
Intelligence objectives can be furthered by correlating information from multiple drives.

- Where any drives were used by the same organization?
- What names/places/email addresses are in common?
- Which drives were used in a place or at a time of interest?
Example problem: Who owned this disk drive?

Approach #1: Find Microsoft Word files; determine owner.

- Needs forensic skill.
- Requires complete documents.

Approach #2: Compute a histogram of all email addresses.

- Works with any file system.
- Works with incomplete data.

The email histogram works even if you can’t find any files.
The email histogram approach works quite well.

Drive #51: Top email addresses (sanitized)

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

(Can we automatically sanitize this kind of information?)
Cross-Drive Forensics systematizes this approach.
“First Order Cross-Drive Forensics” analyzes each drive with a filter.

![Graph showing Unique CCNs versus Total CCNs across different drives.]

Drives with high response warrant further attention.
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 disk #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:

’CHASE NA|5422-4128-3008-3685| pos=13152133
’DISCOVER|6011-0052-8056-4504| pos=13152440
.’GE CARD|4055-9000-0378-1959| pos=13152589
BANK ONE |4332-2213-0038-0832| pos=13152740
.’NORWEST|4829-0000-4102-9233| pos=13153182
’SNB CARD|5419-7213-0101-3624| pos=13153332
With a “credit card number detector,” we can rapidly identify drives with leaked consumer information.
Second-order analysis uses correlation techniques to identify drives of interest.
Second-order analysis uses correlation techniques to identify drives of interest.

In this example, three pairs of drive appear to be correlated.
Let’s look at drives #171 and #172 again.

Cross-drive analysis tells us that #171 and #172 are from the same medical center.

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.
Second-order analysis:

Identifiers:

- CCNs
- Email addresses
- Message-IDs
- sector hashes

Possible Uses:

- Identifying new social networks
- Testing for inclusion in an existing network.
- Measuring dissemination of information
Reactions to this research

Legislative: “Fair and Accurate Credit Transactions Act of 2003”

Technical: Modifications to MacOS & Windows
Looking Forwards

Research Agenda:

• Fix security & privacy in current systems.
• Create clean new systems.
• Use forensic tools to make privacy arguments.
• Make security zero-click.

Pervasive HCI-SEC:

• Use signatures to fight phishing.
• Replace PKI with Key Continuity Management (KCM).
• Secure, privacy-aware data replication.

Questions?