Lessons learned writing digital forensics tools and managing a 30TB digital evidence corpus

Simson L. Garfinkel
Associate Professor, Naval Postgraduate School

August 7, 2012

http://simson.net/

https://domex.nps.edu/deep/

"The views expressed in this presentation are those of the author and do not reflect those of the Department of Defense or the US Government."

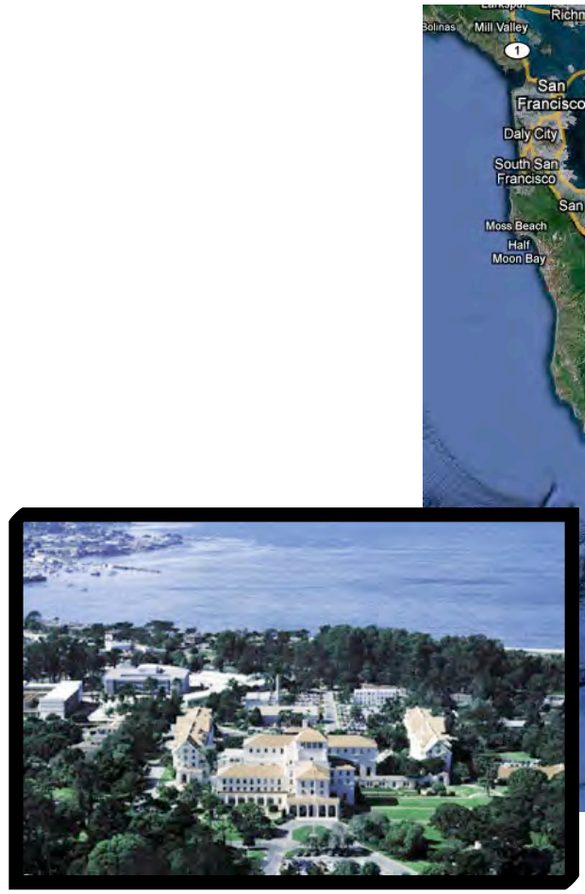
NPS is the Navy's Research University.

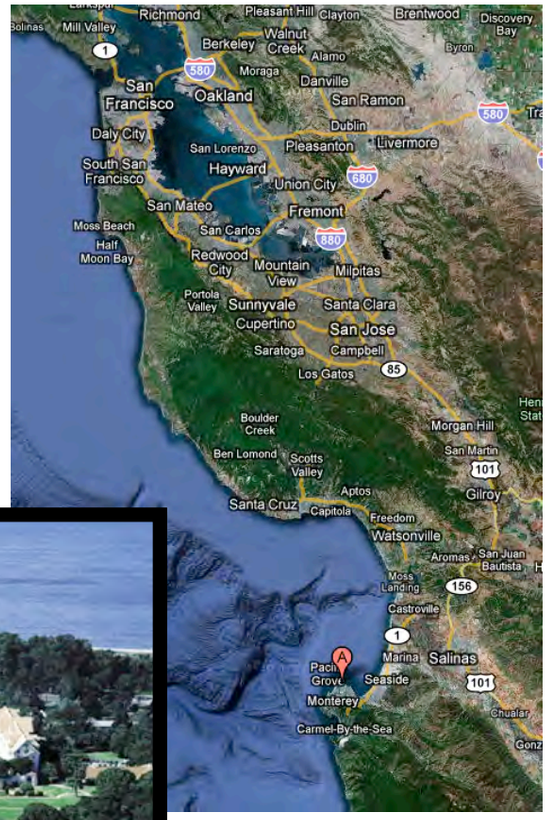
Monterey, CA — 1500 students

- US Military (All 5 services)
- US Civilian (Scholarship for Service & SMART)
- Foreign Military (30 countries)

Schools:

- Business & Public Policy
- Engineering & Applied Sciences
- Operational & Information Sciences
- International Graduate Studies





NCR Initiative — Arlington, VA

- 8 offices on 5th floor, Virginia Tech building
- Current staffing: 4 professors, 2 lab managers, 2 programmers, 4 contractors
- **OPEN SLOTS FOR .GOV PHDs!**

The Digital Evaluation and Exploitation (DEEP) Group: Research in “trusted” systems and exploitation.

“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 and Former Partners:

- Law Enforcement (FBI & Local)
- DHS (HSARPA)
- NSF (Education)
- DoD (JIEDDO & Others)

Traditionally forensics was used for *convictions*.

The goal was establishing possession of *contraband information.*

- Child Pornography
- Stolen documents.
- Hacker tools

Forensics established:

- Data *presence*.
- Data *provenance — where it came from.*

We want to empower forensics for *investigations.*

Extraction — What information does the target have?

- contacts, calendar, documents

Fusion — Putting data together from a single source.

- Timeline generation

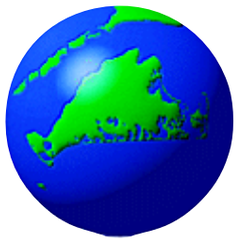
Correlation — data from multiple sources

- Who has the *same information?*
  - — *Identifies members within the organization.*
  - — *Identifies the whole from a part*

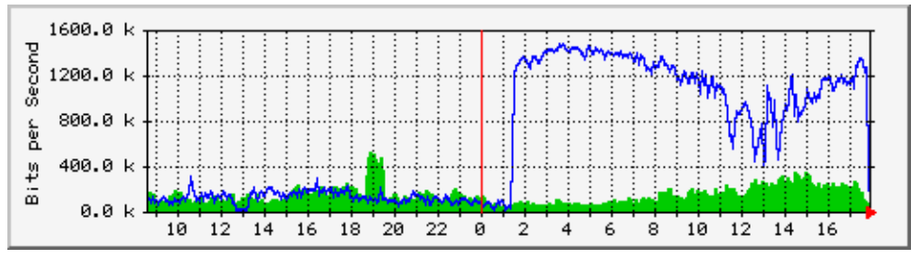
Big goal: automatically identify *actionable information.*

I started working digital forensics in the 1990s

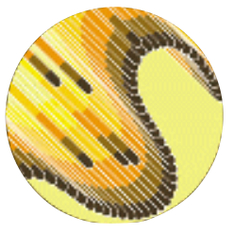
1995 — Vineyard.NET



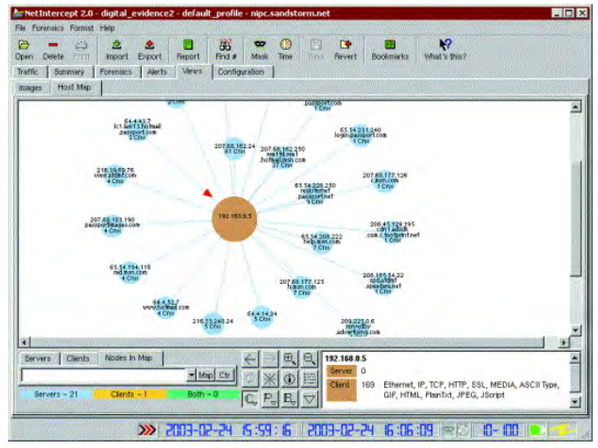
- Used forensics to investigate break-ins
- Limited forensics-related consulting
  - —*We saw forensics as computer security*



1998 — Sandstorm Enterprises



- PhoneSweep — Telephone scanner
- NetIntercept — Network forensics system



1998 — The “drives” project

- Used computers purchased for PhoneSweep had sensitive data
- I started buying drives *for the data*.
  - —*Developed automated analysis techniques for PhD thesis.*

DFRWS 2010:
"Digital Forensics Research, the next 10 years"

Digital Forensics Faces 5 Key Problems:

1. Data size — *more data to process*
2. Encryption & Cloud Computing — *harder to get the data*
3. Mobile phones — *diversity of devices, apps, connectors*
4. RAM and hardware forensics — *RAM changes fast; many places to hide*
5. Tools and training can't keep up — *some devices will never be supported*

Solutions:

- Standard data formats and abstractions (DFXML)
- Standard on APIs (bulk_extractor's scanner API? TSK?)
- Standard data sets for research, education and evaluation (GOVDOCS)
- Emphasize *science*.
- Emphasize correlation — an alternative to the "Visibility, Filter and Report" model

— *http://dfrws.org/2010/proceedings/2010-308.pdf*

This paper looks *backward*, at lessons learned 1995—2012

Why write this paper?

- DF research is hard — but we have have a hard time explaining *why* to outsiders.
- DF is *practitioner* driven — many DF programmers don't consider themselves as such.
- Assistance to funding agencies

What's in this paper:

- Why digital forensics research is different — *it's more than data overload!*
- Lessons learned managing the research corpus
- Lessons learned developing DF tools

What's not in this paper:

- Multimedia issues ("Multimedia forensics is not computer forensics," Böhme et al, '09)
  - *http://pi1.informatik.uni-mannheim.de/filepool/publications/IWCF2009.pdf*

The most important lesson:
Do your best work and help others. You are the professionals.

“The first few meetings were quite tenuous.

We had no official charter.

Most of us were graduate students and we expected that a professional crew would show up eventually to take over the problems we were dealing with.”

—*Reynolds & Postel,*
“RFC 1000: The Request for Comments Reference Guide,”
August 1987

Digital Forensics (Research) is Different

- ✗ blood-and-bullets forensics
- ✗ computer security research
- ✗ reverse engineering

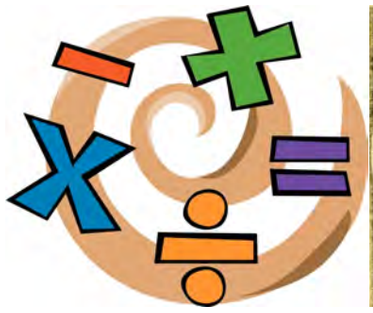
2.1 Diversity is a fundamental challenge of DF

Our charter:

"Analyze any data that might be found on a computer."

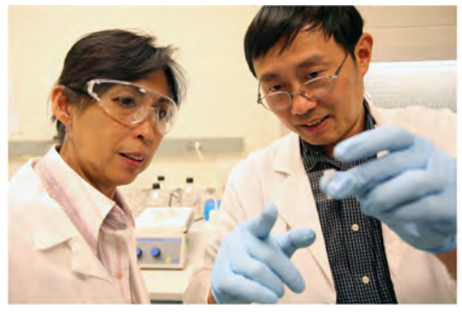
Non-DF research is typically confined to a single area:

energy



math

literature



chemistry

DF must analyze any OS, application, protocol, encryption, etc...

Diversity is more than a multiplicity of file formats...

Data may be *inconsistent* or *incomplete*

- Files that are *deleted* or partially *overwritten*
- Incomplete database records
- Intentionally altered to avoid analysis

Data frequently have no formal specification

- Hacker tools & malware
- Proprietary file formats

We need strategies for systematically addressing diversity

- Exploit similarity and correlation.
  - *Items of interest are frequently repeated.*
- Detect deliberate attempts to hide information
  - *Eliminate the truth and the improbable, and whatever remains must be impossible (and therefore falsified)*
  - *"Improbable" data should be examined for stenography.*

2.2 Data scale is a never ending problem

Scale is continually identified as a DF problem

—*DFRWS 2001:*
"The major item affecting overall performance is data volume: the amount of data collected for analysis of this type is often quite large."

Moore's law scales the targets

- We are using top-of-the-line system to analyze top-of-the-line systems
- We need to analyze in hours (or days) what a subject spent weeks, months or years assembling

∴ We will *never* outpace the performance curve.

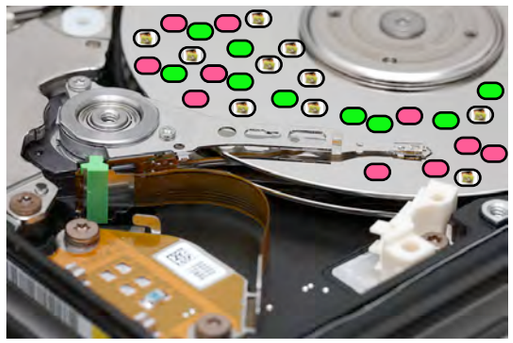
Most "big data" solutions from other fields don't work well with DF

- Budgets — Particle physicists have more $$ per case than we do. (SSC≃1.5PB/month)
- Data diversity — Physics (or even web) data is less diverse than a hard drive data
- Our data fights back — CERN data is not compressed, encrypted, fragmented, or malware
  —*Data complexity dramatically increases I/O and compute requirements*

Use *sampling* and *correlation* to address scale.

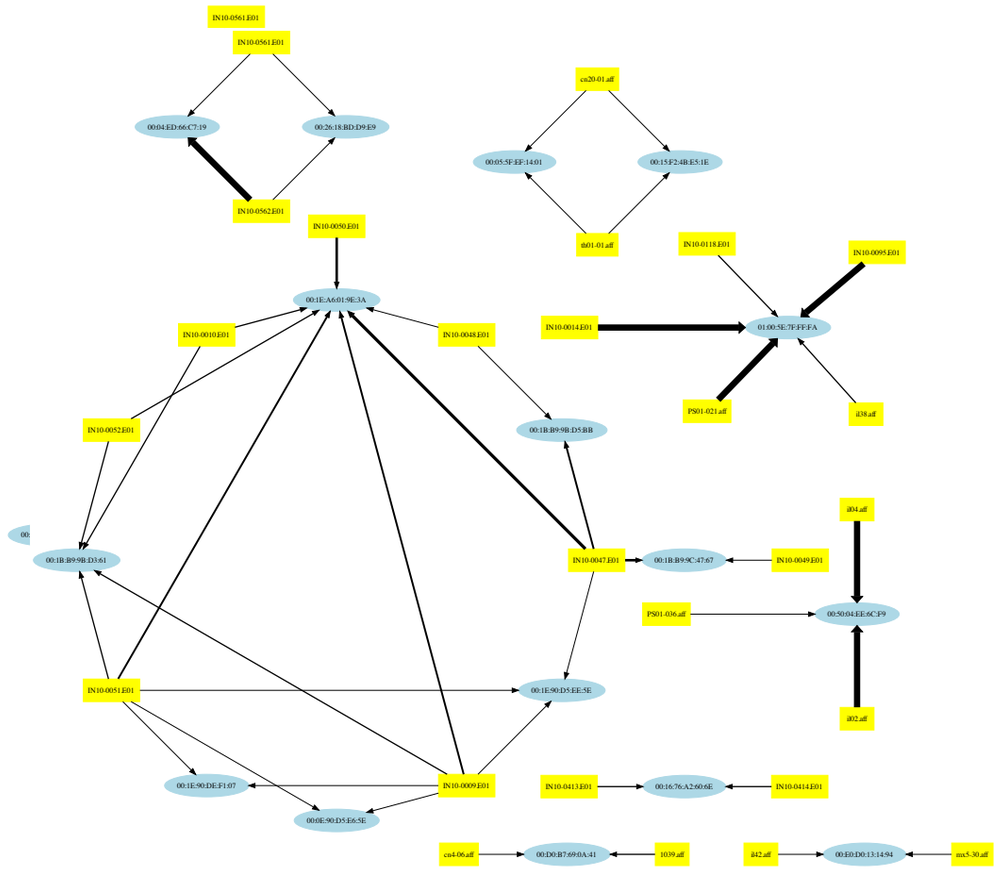
Sampling — “Sub Linear Algorithms”

- Evaluate just as portion of the data; use statistics to draw inferences
- ***Sampling can prove the presence of information!***
- ***Sampling cannot prove the absence of information***
  - ***—just the likely absence***
- “The absence of evidence is not the evidence of absence.”



Correlation:

- Have the data determine what’s important
- Use TF/IDF to remove the mundane (DFRWS 2006)



2.3 Temporal diversity creates a never-ending upgrade cycle

Today's DF tools must process:

- Today's computers / phones / cameras
  - —*Because some criminals like to buy what's new!*
- Yesterday's computers / phones / cameras
  - —*Because criminals are using old devices too!*

Implications for DF users and developers:

- Upgrade DF software as soon as possible.
- DF software will become geometrically more complicated over time....
  - —*... or DF software will adapt on the fly to new data formats and representations.*
  - —*automated code analysis; pattern matching; hidden Markov models; etc.*

2.4 Human capital is bad all over... ... it's especially bad for DF

DF users (examiners, analysts):

- Overwhelmingly in law enforcement.
- Little or no background in CS or IS
- Deadline-driven; over-worked
- Knowledgable users tend to focus in just one particular area.
  - —*Result: It takes two years to train most DF examiners.*

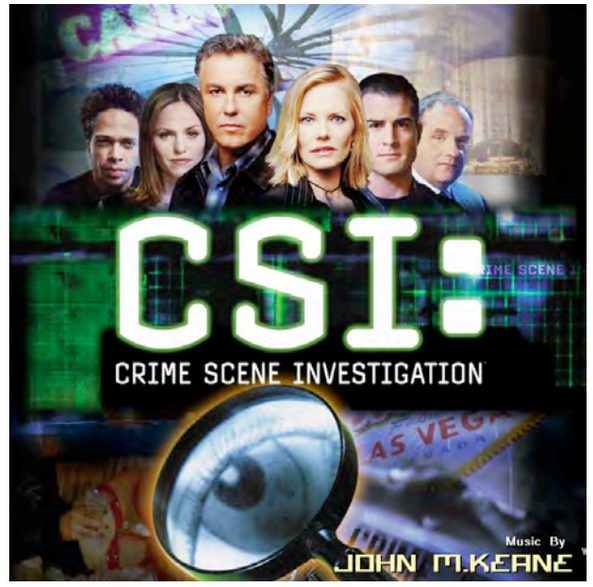
DF developers ("researchers"):

- Data diversity means developers need to know the whole stack
  - —*opcodes & Unicode ⇒ OS & Apps ⇒ networking, encryption, etc.*
- Scale issues means developers need to know HPC:
  - —*threading, systems engineering, supercomputing, etc.*
- Result:
  - —*It's hard to find qualified developers*
  - —*Developers must be generalists*

2.5 The “CSI Effect” causes unrealistic expectations.

On TV:

- Forensics is swift.
- Forensics is certain.
- Human memory is reliable.
- Presentations are highly produced.



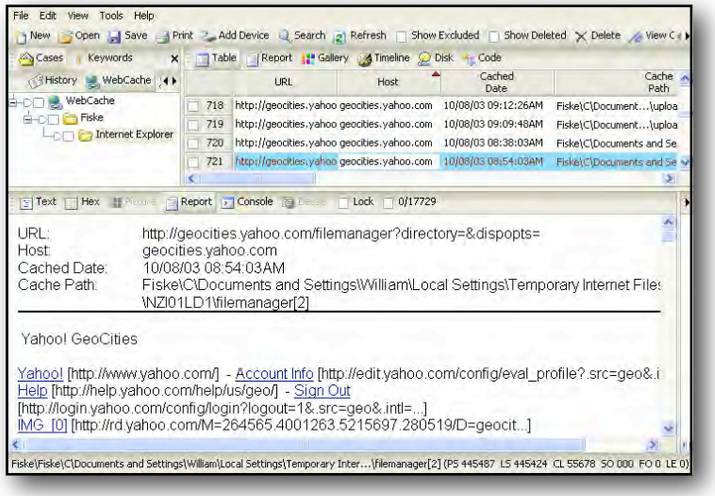
TV digital forensics:

- 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.

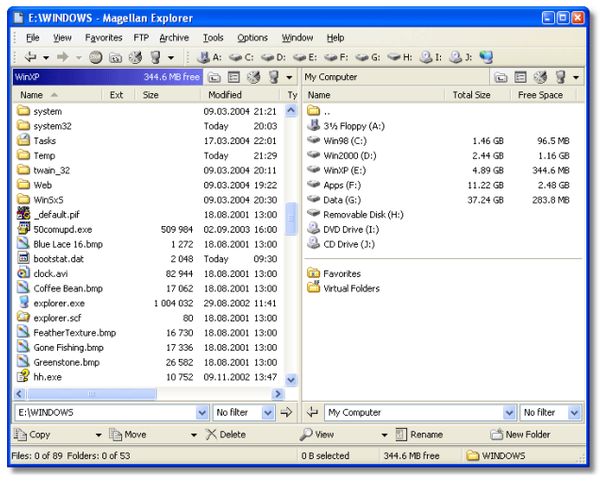
The reality of digital forensics is less exciting.

There are lots of problems:

- Data that is overwritten cannot be recovered
- Encrypted data usually can't be decrypted
- Forensics rarely answers questions or establishes guilt or provides specific information
- Tools crash a lot
- DF tools look a lot like traditional tools



EnCase



Windows Explorer

Result:

— *DF is a difficult process that looks easy*

— *This is not a good place to be*

2.6 Digital Forensics: expensive tools with a limited market

DF tools are expensive to develop:

- Data diversity
- Security critical
- High performance computing

Limited market:

- Consulting firms (more effective tools *decreases* billable hours)
- Police departments (not known for $$)
- Defense (not known for major DF expenditures)

My personal experience:

- It's very hard to stay in business as a tool developer
- Government should have an ongoing role in funding DF research and tool development
- Open source software frequently makes the most sense
  - —*Open Source preserves investment, enables future research, empowers users.*

Real Data Corpus, Harvard University, 2006
800 drives

Lessons learned managing a research corpus

I got started in forensics at Sandstorm Enterprises

In 1998 I bought six used computers for Sandstorm

- For testing PhoneSweep
- All of them had data from the previous owners
- I started buying used drives as a “hobby”

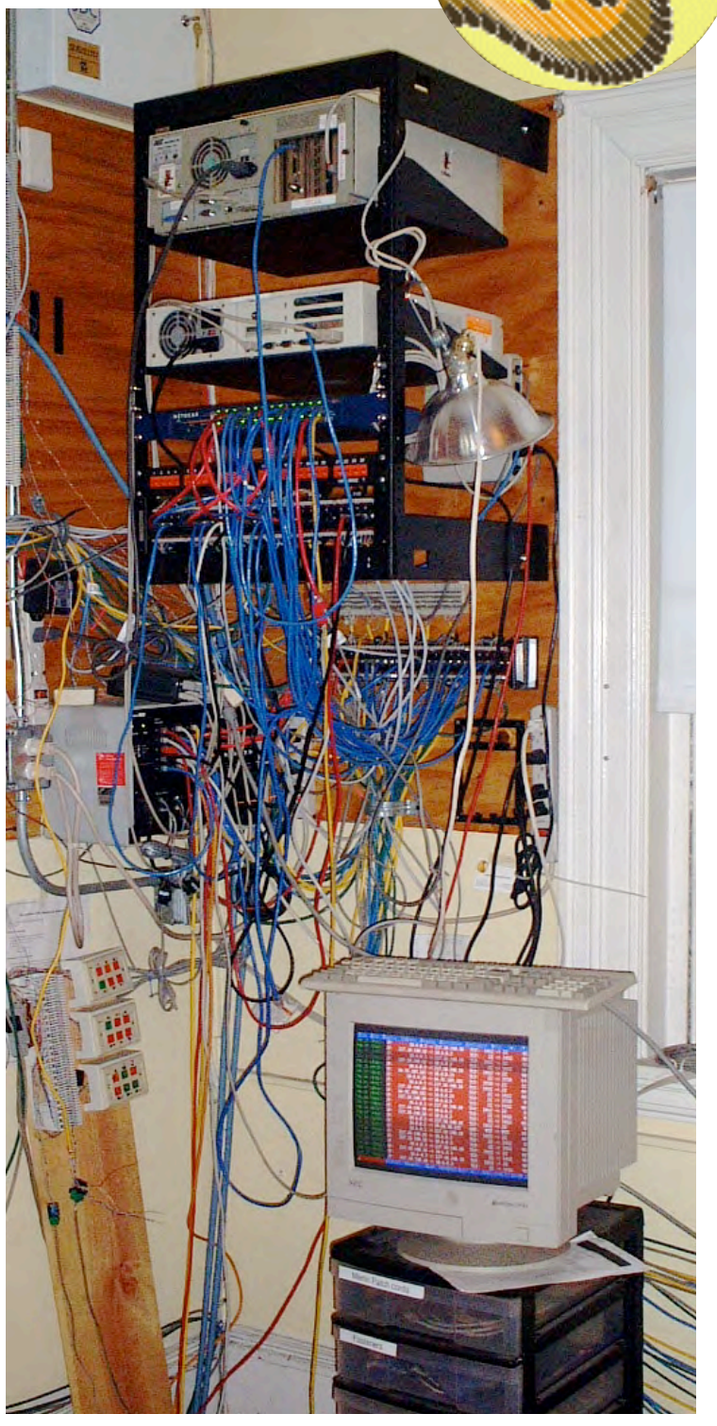
Sandstorm also developed network forensics

- TCP.demux & NetIntercept
- Most customers had no idea what was going over their networks
- Goal was to find the “good stuff”

In 2002 I started the PhD program at MIT CSAIL

- I only had three years of funding
- Made my research looking for the good stuff
- Thesis advisor didn’t want to do forensics...
  - —*So I used forensics to show “intent.”*

CSAIL



I started buying used hard drives as a hobby.

Weird Stuff, Sunnyvale California,January 1999

- 10 GB drive: $19 “tested”
- 500 MB drive: $3 “as is”

Q: “How do you sanitize them?”

A: “We FDISK them!”

In 2002 I started the PhD program at MIT CSAIL

I only had three years of funding

Thesis advisor didn't want to do forensics...

- So I used forensics to show "intent."

There is a thriving used HD market

- Re-used within an organization
- Given to charities
- Sold on eBay

My goal:

- Show how usability failures became security failures.

CSAIL

All Categories [Save this search]

350 items found for **hard drives**

Sort by items: **ending first** | newly listed | lowest priced | highest priced

Picture hide	Item Title	Price	Bids	Time Left
	Lot of hard and floppy drives	$5.50	2	14n
	Lot of hard and floppy drives	$5.50	2	22n
	Lot of hard and floppy drives	$5.50	2	25n
	Lot of 2 hard drives IDE	$8.00	12	29n
	3.2 gig Hard Drives	$180.00	-	59n
	(5) 1.2 hard drives & (15) 10/100 network	$25.00	1	1h 00n
	Lot of 3 Quantum 9.1 gig SCSI Hard Drives	$26.00	6	1h 25n
	IDE HARD DRIVES (3)	$6.50	6	1h 46n
	LOT OF 5 Hard Drives! 3.2 Gig Western Digital	$120.00 $124.95 Buy It Now	-	1h 50n
	QTY 3....IDE Hard Drives 2.5 Gig	$20.50	5	2h 02n
	5 WESTERN DIGITAL 2.5 GIG HARD DRIVES	$30.00	4	2h 03n
	QTY 3....IDE Hard Drives 1.0 Gig	$9.99	1	2h 04n
	Western Digital 850 meg IDE Hard Drives dutch	$6.00	1	2h 57n
	WINDOWS	$6.00	-	3h 18n

PREMIER ISSUE

IEEE

SECURITY & PRIVACY

Building Confidence in a Networked World

JANUARY/FEBRUARY 2003
VOLUME 1, NUMBER 1

Fact over Fiction

Second-Hand Secrets p. 17

IEEE COMPUTER SOCIETY http://computer.org/

Some of the reactions were confused...

"Good luck removing data from this."

"Our prognosis: drive slagging is a fool-proof method to prevent data recovery."

It's *easy* to remove data from a hard drive.

— *You just have to do it.*

3.2 Corpus management: Technical Issues

I imaged 1000+ drives in 1 year.

I stored images as raw.gz

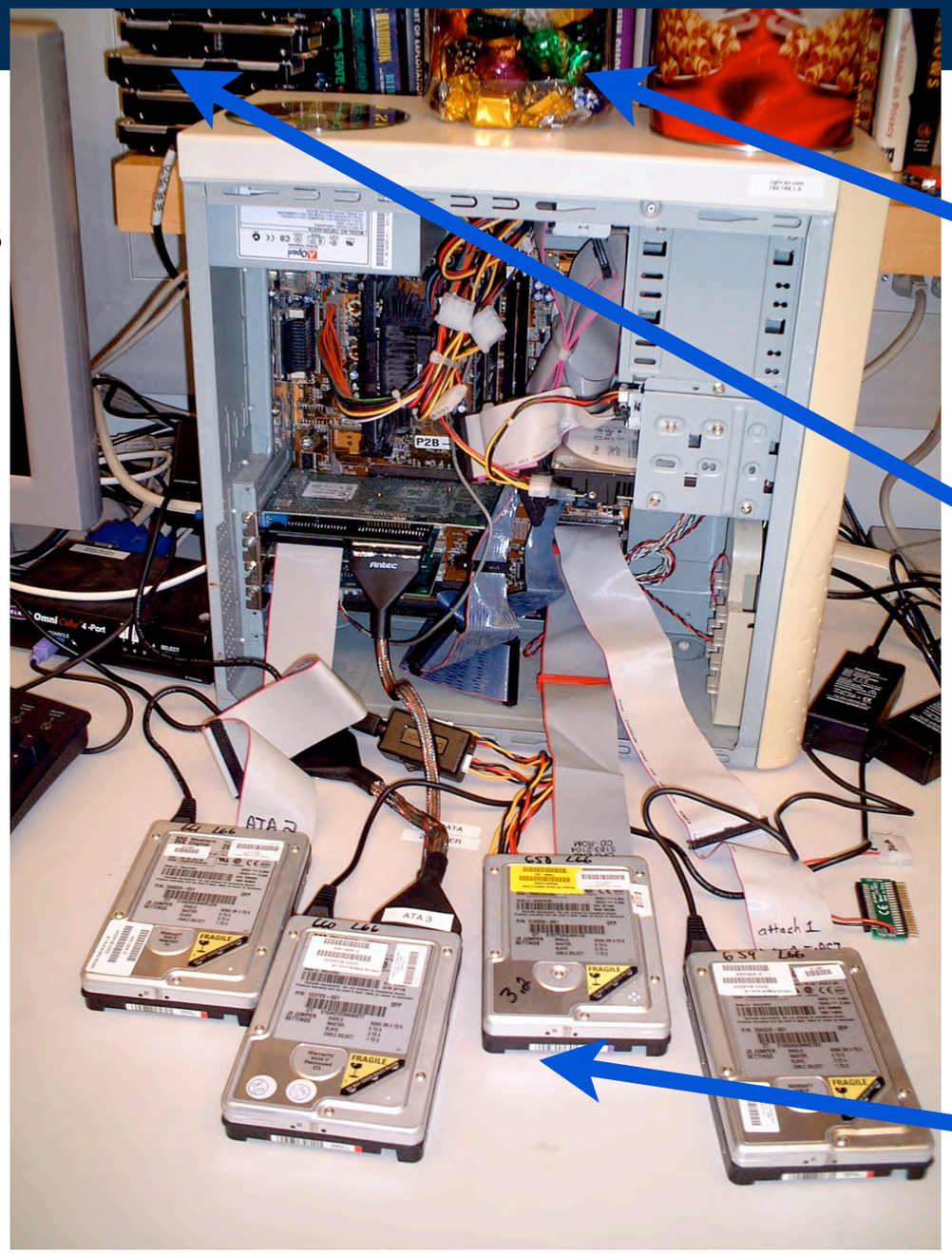
- naturally led me to stream-based forensics

Lessons learned:

- ATA is hot-swappable
- Don't maintain software (aimage) that does the same thing as other open source software (guyimager)

Critical technology:

- Handling read errors
- Automated metadata collection



Drive imaging workstation, Harvard University 2005

Using disk images for *research* required storing all of the disk images *online* for easy access.

I needed to store a lot of images in a small area.

640GB RAID Array (2003)

1.5TB RAID Array (2006)

I wanted:

- simplicity — a single file with all metadata embedded
- convenience — small file names with short paths (ease of use)
- permanence — file names and path names that wouldn't change

Evidence file formats is no longer a research topic.

In 2005 I started on AFF (Advanced Forensics Format):

- Store metadata & data together
- Extensible
- Read & Write, but optimized for archiving
- Support for digital signatures, encryption, chain-of-custody
- aimage imager that did “sparse imaging” and error recovery

Automation is key; any process that involves manual record keeping is going to introduce inaccuracies that will be hard to detect and correct.
Useful data will outlive the system in which it was stored, so make provisions to move the data when you design the system.

Evidence file formats is no longer a research topic.

Since 2010 I have largely given up on AFF:

- E01 can now handle terabyte-sized HDs in a single file
- Joachim Metz's ewfacquire & libewf do an excellent job supporting E01
- EX01 has encryption (and we should soon have an open source EX01 implementation)
- AFF4 is an interesting platform, but it's unlikely to be adopted by EnCase, so I can't use it.
  - —*Most important: avoid replicating other people's work if you can*

Avoid developing new file formats whenever possible.

Kill your darlings.

Do research, and only maintain software that implements a particular function when no other software is available.

Bad ATA drives crash Linux & FreeBSD

Crashes look like wild memory writes.

- ATA spec allows DMA to system memory
- Motherboards probably don't defend against wild DMA.

Question:

- Can we use this as a memory acquisition technique?
- There is "legacy ATA" on many motherboards.

many technical options remain unexplored.

Many bad drives had sensitive data!

- Always read to the "end" of the drive
- Read all the drives in the RAID set

Drives with some bad sectors invariably have more sensitive information on them than drives that were in working condition when they were decommissioned.

3.2 — Corpus Management — File Types and Path Names

Many different modalities:

- Disk images — “drives”
- Memory Images — “ram”
- Scenarios — symlinks to souce images
- Packet Dumps — “net”
- Files — “files”

Many different sources and distribution restrictions:

- Used purchased inside US — “US” (not used by USG)
- Used purchased outside US — “NUS”
- Created by NPS, redistributable — “NPS”
- Created by NIST — “NIST”

Although it is advantageous to have names that contain no semantic content, it is significantly easier to work with names that have some semantic meaning.

Consistent naming scheme on every machine:

/corp/source/modality/description/daughter-files

/corp/nist/rds/rds328/
/corp/nps/files/govdocs1m/123/123456.jpg
/corp/nus/drives/in/IN10-0249/IN10-0249.E01, IN10-0249.E01.txt

Lessons learned with naming

/corp/source/modality/description/daughter-files

/corp/nist/rds/rds328/
/corp/nps/files/govdocs1m/123/123456.jpg
/corp/nus/drives/in/IN10-0249/IN10-0249.E01, IN10-0249.E01.txt

Every data object should have a unique file name.

- Put something *very descriptive* in the file name
  - *Source country*
  - *Scenario name*
- Don't change file names.
  - *If you must change names, try to have the old name inside the new name*
    ubnist1.E01 -> nps-2009-ubnist1.E01
- It's okay to change directories.

Different users want different subsets of the corpus...

- It's best if they use the same file hierarchy.

Place access-control information as near to the root of a path name as possible.

Anti-virus and indexing cause numerous problems

Disable AV and indexing on your corpus.

- Forensic data has viruses
- Corrupt and unstructured data frequently crash indexers

Unfortunately, exceptions need to be reapplied:

- After software updates
- After OS upgrades
- When new external HDs are attached.

Configure anti-virus scanners and other indexing tools (e.g., Apple's build_hd_index) to ignore directories that might contain raw forensic data.

3.2.9 There is no good way to distribute a 30TB data set

Approaches we have tried:

- Transferring over the Internet by scp, rsync, BitTorrent, uftp, Aspera
- Sending 2TB internal SATA drives
  - *Need SATA dock.*
  - *File System Choice: ext2/3/4? HFS? NTFS?*
  - *(NTFS seems best choice for read-only)*

- Sending 8 drives and a DroboElite
  - *Overcomes single-drive failure*
  - *VERY SLOW!*

Solutions developed by other disciplines for distributing large files rarely work well when applied to DF without substantial reworking.

Added complications — “bit rot” long term storage — off track writes

- *Evaluating the Impact of Undetected Disk Errors in RAID Systems* *https://www.perform.csl.illinois.edu/Papers/USAN_papers/09ROZ01.pdf*
- *Modeling the Fault Tolerance Consequences of Deduplication* *https://www.perform.csl.illinois.edu/Papers/USAN_papers/11ROZ02.pdf*

3.3 Corpus management — Policy Issues

3.3.1 — Privacy Issues

3.3.2 — Illegal content — financial, passwords and copyright

3.3.3 — Illegal content — pornography

3.3.4 — Institutional Review Boards

Even if something is legal, you may wish to think twice before you do it.

3.3.1 — Privacy Issues

Information in the RDC is not legally “private”

- “The reasonableness of a search for Fourth Amendment purposes ... turns upon the understanding of society as a whole that certain areas deserve the most scrupulous protection from government invasion. There is no such understanding with respect to garbage left for collection at the side of a public street.”

  —*CALIFORNIA v. GREENWOOD, 486 U.S. 35 (1988)*

- In practice, we avoid disclosing PII because that’s the right thing to do.

Copyright on user-generated material

- Users do not transfer copyright to us, but we do have some rights in our copy
- “First Sale” doctrine — “In our view, the copyright statutes, while protecting the owner of the copyright in his right to multiply and sell his production, do not create the right to impose, by notice, such as is disclosed in this case, a limitation at which the book shall be sold at retail by future purchasers, with whom there is no privity of contract.”

  —*Bobbs-Merrill Co. v. Straus, 210 U.S. 339 (1908)*

- “Fair Use” — four part test. 1) purpose of the use (non-profit educational); 2) nature of the copyrighted work; 3) The amount of the work that is copied; 4) the impact of the use on the market value for the copyrighted work.

3.3.2 & 3.3.3 — Illegal Content of all types

"The Counterfeit Access Device and Computer Fraud and Abuse Act"

- Passed by Congress in 1984
- Outlaws possession of "access devices" with intent to commit fraud.
- Financial information (credit card numbers) and passwords are access devices.
- The key issue is intent — I don't have the intent do defraud.

Copyright — rely on "Fair Use" (17 USC §107)

- Four part test. 1) purpose of the use (non-profit educational); 2) nature of the copyrighted work; 3) The amount of the work that is copied; 4) the impact of the use on the market value for the copyrighted work.
- The RDC doesn't impact the value of the data, and it's non-profit.

Conventional pornography

- The RDC has ***lots*** of pornography in it
- No access given to minors

Obscenity (e.g. child pornography)

- We can't determine if something is really child porn...
- When we find information "suggestive" of child pornography, we remove it and notify others.

3.3.4 Institutional Review Board issues

In the United States, federally funded research involving human subjects *must* be reviewed by an accredited Institutional Review Board.

Note:

- "Human subject" means a living individual about whom an investigator conducting research obtains:
  - —*Data through intervention or interactions with the individual, or*
  - —*Identifiable private information*
- "Research" meansd "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge."

Options:

- Make the RDC public (so it's not private information)
  - —*That would be unethical.*
- Don't do "Research"
- Get IRB approval (and that's what we do)

Lessons learned developing DF tools

4.1 Platform and Language

DF platform:

- Windows is the dominant platform for users.
- Mac & Linux seem to be the dominant platform for researchers & developers.
- We need to support all platforms.

4.1 Platform and Language

Requirements for a DF language:

- Good for string operations & handling large blocks of data
- Good for interpreting binary data in different formats
- Good support for Unicode
- Compiled code should be *fast ;* Should be safe

	String operations	Binary operations	Unicode	Safety
C		✔		
C++	✔	✔	fair	fair
Python	✔		✔	✔
Perl	✔		✔	✔
C#	✔	✔	✔	✔
Java	✔	✔	✔	✔

— Java seems to run faster than C/C++ in some situations.

— C# should be the language of choice, but support on Linux & Mac is problematic.

4.2 Parallelism and high performance computing

Clearly, we need to investigate parallelism and HPC.

Exotic Hardware: CellBE & GPU

- Our MD5, SHA1 & AES implementations on CellBE ran 1-6 Gbytes/sec
  - —*But IBM failed to deliver a 10GigE blade that worked with Cell Broadband Engine.*
- GPU systems are *great* for parallel execution.
  - —*But you need multiple passes on the same data; GPU I/O isn't fast enough for hashing...*

bulk_extractor uses parallel threads processing different data pages

- Easy-to-implement, but limited usefulness.

We have a cluster with 2000+ cores

- We have had a LOT of problems making forensics work on the cluster.
- Most problems result from big data and large I/O requirements.

4.3 All-in-one tools vs. single-use tools

Different goals:

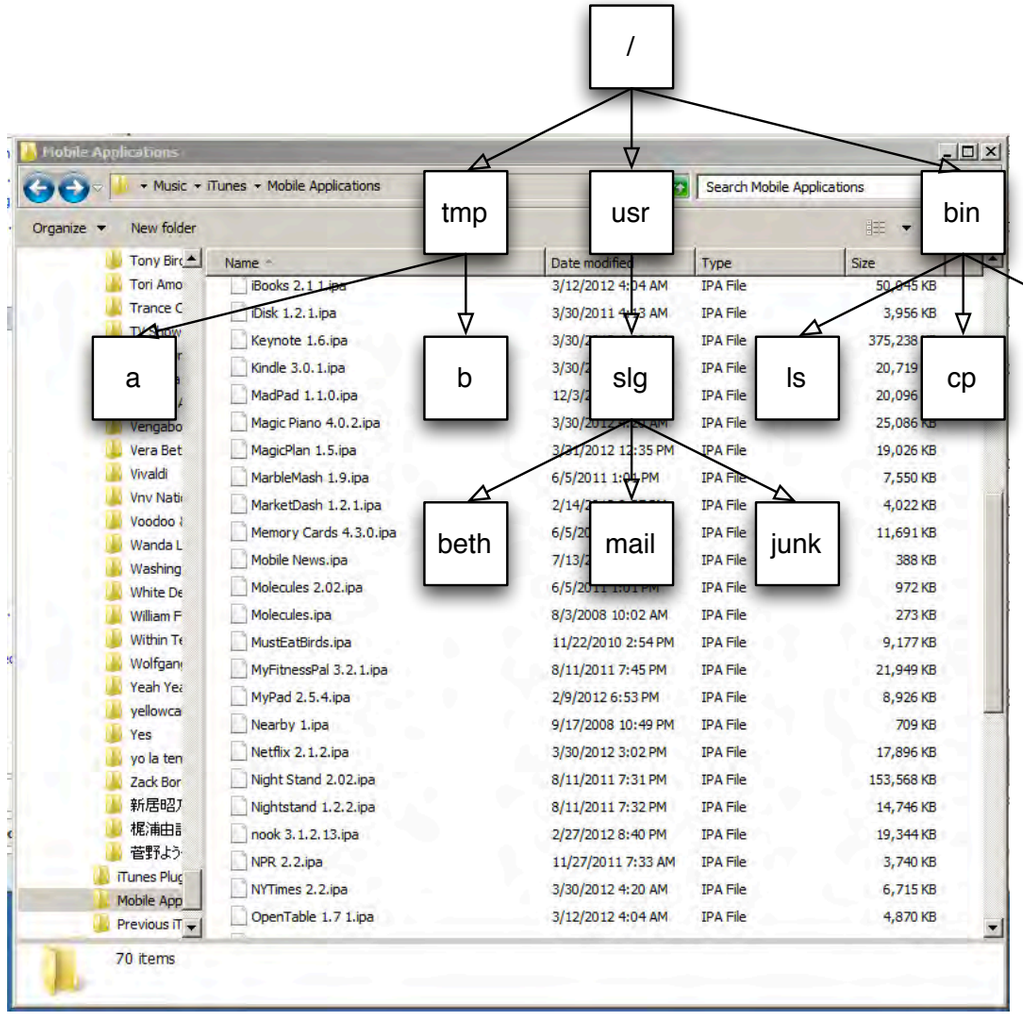
- Examiners want all-in-one tools.
- It's easier to create single-purpose tools.

Much of what tools do is redundant:

- data ingest
- decoding and enumerating dta structures
- preparing final report.

Suggestion:

- Have tools generate multiple outputs.
  - —*Easy-to-read for human analysis*
  - —*Complete data for machine analysis*
- Try embedding XML in PDFs.



4.4 Evidence container file formats

Forensic tools need to be able to handle:

- raw; split-raw; EnCase; AFF (*)
- packets (libpcap)

Options for disk images:

- Use SleuthKit to read all of the files.
- bulk_extractor has a C++ iterator that provides access to data and metadata

Options for packets:

- libpcap (libpcap-dev must be installed)
- “pcap_fake” (currently under development for Windows version of tcpflow)

4.5 DFXML metadata and provenance

There's lots of structured data to represent:

- File names, locations, MAC times, etc.
- Which program processed the data:
  - *Which version; where compiled, compiler flags, etc.*
  - *Where it was run, how long it took, etc.*

Originally programs kept this information in many different places:

- SleuthKit "body" file; Log files; Etc.

DFXML is a single, unified way of keeping all of this information.

- Arose out of personal need.
- Decided that it would be better not to reinvent a storage format.
- XML has broader support than other formats, more people know it, and it can deal with GB-sized data.
- Easiest way to get support for DFXML: add it to open source programs.
- Working now to merge DFXML with CyBox

Related Work

Walls, Levine, Liberatore & Shields (2011b) "Effective digital forensics research is investigator-centric"

"Experiences like ours in deploying well-used tools based on novel forensic research are rare. More typically, computer security research aimed towards forensic applications ahve little or not impact—often because the researchers are poorly acquainted with the real-world problems faced by forensic investigators and the constraintes placed on solving them."

- Digital forensics is investigator-centric, and unless developed with an understanding of the restrictions that investigators are under, most novel results cannot and will not be adapted.
- The value of a new technique depends in part on its complexity and therefore it must be judged against simpler options available to investigators.
- Forensic techniques are most valuable when addressing the cmost common adversary, not the strongest.
- Forensic investigations seek to find the person responsbile rather than stopping at a machine or line of code.

—*http://prisms.cs.umass.edu/brian/pubs/Walls.hotsec.2011.pdf*

In conclusion...