Introduction to Digital Forensics Research

Dr. Robert F. Erbacher
Utah State University
Department of Computer Science
Logan, UT 84322

Table of Contents

• Introduction/background
• Forensic processes
  – System preservation
  – Analysis
• Lackings and Research Directions
• Computer Forensics
• Network Forensics
• Legal issues
  – Example with data validation
• Conclusions

Introduction/Background

Forensic Goals

• Analysis of an intrusion
• Analysis of digital storage devices to locate evidence
• Legal admissibility
  – Must be considered in criminal cases
  – Must be considered in non-criminal cases, i.e. employee demotion/termination

Forensic Goals

• If legal admissibility is NOT considered, is it forensics or just analytics?
  Locating how an intrusion occurred to remove the vulnerability
• Must consider how the analysis process or technique might be legally challenged
  – Analytics++
• Must consider how criminals can compromise the validity of the data/technique

Definitions

• American Heritage Dictionary
  – The art or study of formal debate; argumentation.
  – The use of science and technology to investigate and establish facts in criminal or civil courts of law.
• Dictionary.com Unabridged
  – pertaining to, connected with, or used in courts of law or public discussion and debate.
  – adapted or suited to argumentation: rhetorical.
  – the art or study of argumentation and formal debate.
Goals

• Promote good forensics research
• Disseminate what forensics really requires
• Must consider how the technique/process can be challenged/compromised anywhere in the process
  – Challenged by expert witnesses
  – Compromised by criminals
  – Compromise can be proposed by expert witness
• Must identify what forensics means and stop inappropriate usage of the term in order to maintain its integrity
  – Don’t inhibit needed research

Focal Directions

• Computer Forensics
• Network Forensics
• Data Forensics
  – Forensic accounting
  – Database forensics

Digital Forensics

• Computer Forensics: The analysis of computers and hard drives from crime scene, etc.
  – Developing interaction and visualization techniques for the analysis of computer hard drives
  – Developing techniques for identifying and collecting data hidden in unusual hardware locations.
  – Developing techniques for collecting and analyzing data from mobile devices
  – Evaluation and comparison of available techniques

Computer Forensics

• The analysis of computers and hard drives from crime scene, etc.
• Research focuses on
  – Locating hidden information
  – Handling large amounts of organized data

Digital Forensics

• Network Forensics: The application of forensic processes to networks of interconnected devices. Mainly relates to systems compromised from a network based intrusion.
  – Techniques for forensically collecting applicable data. How can this data be validated for forensic use, i.e., such the it can be used in criminal proceedings
  – Analysis of attacks in order to determine what data needs to be collected in order for said attacks to be collected. The ultimate goal is to allow for the more efficient collection of networked data
  – Techniques for the analysis of very large scale data associated with networks of computers and mobile devices. This will include network traffic data, system log data, router data, DNS data, etc

Network Forensics

• The application of forensic processes to networks of interconnected devices
  – Mainly relates to systems compromised from a network based intrusion
• Research focuses on
  – Locating hidden information
  – Handling large volumes of chaotic data
Data Forensics

- Who created the data?
- Who accessed the data?
  - VA laptop
- Validate the integrity of the data
  - Has it been changed?
  - How reliable is it?
- Especially related to databases
- Accounting forensics

Forensic Processes

Recovery

- Prosecution or analysis not a concern
  - No dependence on data
  - Often the case for many businesses
  - Non-critical systems
  - No important data compromise
- Restore from mirror or backup
- Rebuild system

Research

- Data is critical but collection can be adhoc
- Delay in recovery likely acceptable
- Goal is to understand how compromise was done
- Data copied from drive to repository
- System recovered in a reasonable fashion

Prosecution

- Typical forensics, ala CSI
- Must follow exact procedures and protocol
  - This is evidence
  - System must be preserved
  - Chain of custody must be shown
  - Securely stored
  - Document analysis process
  - Follow accepted/proven analysis processes
Basic Forensic Process

Forensic Processes for Research

Other Considerations
- Cognitive task analysis (CTA)
  - Done for network analysis
  - Deviations?
- Human perception
  - What improves/detracts visual comprehension
- Scalability
  - Need to enable analysis of terabytes of data
- Current techniques result in data needing further analysis

General Workflow Pattern

Task Flow Diagram

System Preservation
**System Preservation**

- Hard drive evidence never touched directly
- Special hardware used to make images
- Only images are analyzed
- May not know what type of file system was used which requires additional recovery
- Original hard drive stored securely
  - Chain of custody logs maintained
  - Document all handling of drive

**System Preservation Issues**

- Password protected drives
  - Will only boot while connected to original controller
- Computer networks
  - What systems need to be preserved?
  - An attack could touch many systems
  - Only preserving compromised system may lose data
  - Storing all systems isn’t feasible
- Raid devices?

**System Preservation Steps**

- **Microsoft windows**
  - Photograph screen
  - Annotate any programs running
  - Pull power cord from wall
- **UNIX**
  - Photograph screen
  - Annotate any programs running
  - If root: Sync; sync; halt
  - Otherwise: Pull power cord from wall

**Analysis**

- Analysis requires examination of available data and locating relevant information
- In both cases enormous volumes of data
- Typical techniques rely on command line
- Newer techniques rely on visualization
  - Various definition of visualization

**Lackings and Research Directions**
Current Lackings

• What does data duplication mean?
  – Encryption/attribution/validation

• What techniques are needed?
  – i.e., what capabilities do analysts need

• What is required for legal validity?
  – Conversely, what will prevent legal validity

Research Directions

• Preservation of new device types
  – Mobile and embedded devices
  – Without system modification/impact
  – Encrypted data/drives

• Pre-emptively handle anti-forensics

• Formally verify techniques

• Identify what data must be collected

Research Directions

• New techniques to analyze data
  – Scalability
  – Anti-forensics
  – Obfuscation

• Fulfill due process
  – Where does the IRB fit in

• Education

• Legal admissibility

Computer Forensics

Scale of the Problem

• Size of hard drives
  – 750GB each
  – 1+TB servers becoming common

• Many more desktops

• More obfuscated data

Command Line Based Analysis

• Find, Sed, Grep
  • Ls, Cat, Strings
  • Stat, File
  • Less, head, tail
  • Sort, uniq, diff
  • Script, date, uname –a
  • Netstat, nc
  • Isof, uptime, who, ps, top, fuser, strace
  • Customized versions of commands often used
  • The options make the commands useful
Statistical Approaches

- Currently under investigation
- Can we identify statistically that a file is anomalous
  - That it contains hidden information
- Create statistical curve of file
  - Avg. values per window
  - Does the curve uniquely identify file types?
  - Other statistical parameters

Visualization Techniques

- Simple techniques
  - Graphs
  - Charts
  - Explorer
  - Don’t handle the volume of data
  - Poor exploration and analysis
- Advanced visualization techniques
  - More abstract representations
  - Disk defragmenters

Analyzing Disks

- Many file attributes of interest
- Many attributes are file system dependent
- Specific scenarios will identify additional attributes of importance
- Owner sophistication identifies attributes of importance
- Goal is to find traces of programs that were run.
  - ID fact that a wiper was run
  - FBI can recover from 20+ wipes. ID what was wiped

Layers of Analysis

- Physical storage media analysis
  - Memory analysis
- Volume analysis
  - Swap space analysis
  - Database analysis
- File system analysis
- Application/OS analysis

Attributes of Interest

- File type
  - Especially mismatches
- Access and modification times
- File size
- File location
- File contents
  - Especially key fragments of information

Computed Attributes of Interest

- May or may not be available
  - Journal file system
  - Word processor
  - Number of times accessed
  - Frequency of access
  - Duration of accesses
  - Frequency of changes
  - Changes in file size
  - Changes in file type
  - Relocation of the file
  - Many others
Data Obfuscation

- Rename
- Move
- Delete
- Encrypt
- Compress
- HPA
- DCO

Ways of Hiding Data

- ADS
- Serialization
- Steganography
- Append to file
- File slack space
- Store on video card
- Store in system bios

- Putting data in hidden partition
- Deleting the data partition
- Storing data in bad clusters
- Storing data in IO buffer
- Storing data on NAT/NAS
- Covert Channels
- Storing data on removable disks
- Storing data on open file shares
  Can we back track this?

Where is Waldo?
errrr, Stalker…

Where is Tiger?

Why is the Laptop Failing?
Where Did the Attacker Go?

Visual Computer Forensics

- Visually represent important attributes
- Allows for rapid identification of anomalies
  - List of files resulting from search vs.
  - Picture of files matching with attribute values shown
- Block view – disk defragmenter
- Treemap view – space filling graph

Block Diagram View

Block Diagram View 2

Block Diagram View 3

Treemap View Example 1

From “Treemap Visualizations of Newsgroups” by Fiore and Smith
Treemap View Example 2

Treemaps 1

Treemap View Example 2

From “Treemap Visualizations of Newsgroups” by Fiore and Smith

Treemaps 1

Analyzing Network Data

• Network traffic data
• System log files
• Command history
• Possibly reverts to computer forensics
• Correlate necessary data

Network Forensics

Data Collection Issues

• What systems on a network are evidence?
  – A single comp broken into
  – Potentially evidence on many
• Syslog services
  – Can be tampered with in transit
  – Secure syslog, how admissible is it?
  – How validatable?

Temporal Data

• How validatable is time
• How can time be attacked
• Slight time drift can change order of events
  – Can change meaning of event sequence
  – What scale of temporal activity is challengeable
• What about dropped packets
  – For network traffic data

Analyzing Network Data

• What systems on a network are evidence?
• Syslog services
• Command history
• Possibly reverts to computer forensics
• Correlate necessary data

**System Log Files**

- Touched network systems may have evidence in system log files
- Goal is to collect these log files for evidence without the need for the whole system
- Send log files to central repository
- Tag log entries to validate
  - Sufficient capability not yet available

**Attributes of Interest**

- Src/Dst IP’s/Ports
- Packet time
- Connection type
- Active/inactive flags
- Payload
- Checksum validity
- Packet length

**Computed Attributes of Interest**

- Number of packets sent
- Frequency of packets sent
- Number of different systems sent to
- Total and average bandwidth used
- Duration of connections
- Number of connections

**Network Forensic Visualization**

- Show important attributes of network data
- Reduce clutter and obfuscation
- Allow for data reduction and analysis
- Design for task at hand
  - Low and slow scan detection/analysis

**Identification Issues**

- Lack of trace back
  - Unmanaged switches
  - Hubs
- Delay in trace back
  - DHCP
  - Wireless
- Any trace back may be challengeable?

**Data Collation Example**
**Goals**

- Legal applicability
  - Extract potentially relevant data without examination of irrelevant data
  - Child porn hidden in financial data
  - Definitive, validatable algorithm
  - Legally admissible data collection
- Legal admissibility

**Daubert Challenges**

- Radar speedgun
- Breathalizers
- FBI fingerprint DB/matching algs
  - Prove someone isn’t included incorrectly
  - Prove someone isn’t excluded incorrectly
- Increasing sophistication by lawyers of digital evidence

**Federal Rules of Evidence Applied to Digital Data**

- Us laws
- Hearsay
  - Challenge expert witness
- Validate science behind data
  - Processes
  - Algorithms
  - Raw data generation
- With advanced techniques how do you validate to jurors that the algorithm is correct?
- Error rates, proving validity of data
**Analyst Process Consistency**
- Individual groups have different processes
- Each analyst has a different process
- A analyst will continuously see new scenarios and may not react appropriately

**Legal Admissibility of Process**
- What process is appropriate for analysts
  - Must ensure the legal admissibility of the evidence
  - Once process is set, can it be countered
  - What if process is attacked?
- How can analysts be trained in the process
  - How can we avoid inadmissible deviations

**Legal Ramifications**
- What is an admissible analysis process?
  - Analysts use their own process
- What happens with the lack of process consistency?
- What is the definition of in plain sight?
  - When dealing with a hard drive
- Who was at the computer?

**Validation Mechanisms**
- Algorithm analysis
- Formal methods
- Mathematical analysis
- Testing
- Peer evaluation
- External evaluation
- What guarantees the best level of accuracy
  - Electronic voting machines

**Who Will Defend Technology**
- Scientist/Researcher
  - The creator
  - Are all scientists suitable as expert witnesses?
- User
  - The analyst
  - Do analysts have sufficient knowledge of the techniques?

**Who Will Defend Technology**
- Attorney
  - We killed them all yesterday
- Third Party
  - Not as effective as the creator
  - Validate against known tools
- What about open source software?
Teachability

- Will the jury comprehend the validation?
- Will the jury believe the validation?
- How attackable is the validation?
- How defensible is the validation?
- Being a good expert witness is a whole issue in and of itself. Whole books on the subject.

Example with Data Validation

Mapping Casey's Certainty Levels to Syslogs

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
<th>Relationship to Syslog</th>
</tr>
</thead>
<tbody>
<tr>
<td>C9</td>
<td>Erroneous/Highly Uncertain</td>
<td>Programmatic errors while coding the syslog/tdslogd protocol. [17]. An attack occurs by exploiting this vulnerability.</td>
</tr>
<tr>
<td>C8</td>
<td>Erroneous/Highly Uncertain</td>
<td>A faulty syslog file with manipulated entries.</td>
</tr>
<tr>
<td>C7</td>
<td>Erroneous/Uncertain</td>
<td>In the event of an attack, the only evidence that is available is the organizational syslog file. Distributed evidence preservation—this paper—not attempted.</td>
</tr>
<tr>
<td>C6</td>
<td>Probable</td>
<td>Syslog variants, namely, Syslog-sign and Syslog-fault, have this level of certainty.</td>
</tr>
<tr>
<td>C5</td>
<td>Probable</td>
<td>Syslogs and authentication traces that are stored and transmitted in plain text can be classified to have this level of certainty.</td>
</tr>
<tr>
<td>C4</td>
<td>Almost Certain</td>
<td>This level of certainty specifies evidence to be tamperproof and asserts a match between independent sources of evidence, which in this case are the authentication traces and sylogs. The evidence at this level, however, can be erroneous due to temporal loss or data loss. The proposed method belongs to this level of certainty.</td>
</tr>
<tr>
<td>C3</td>
<td>Certain</td>
<td>If authentication traces were validated at every system that they were generated on and more importantly, at intermediate stages in the routing to the syslog server, syslog evidence would then have this level of certainty.</td>
</tr>
<tr>
<td>C2</td>
<td>Certain</td>
<td>A patchy syslog file with manipulated entries.</td>
</tr>
<tr>
<td>C1</td>
<td>Certain</td>
<td></td>
</tr>
</tbody>
</table>

Sequence Diagram of Proposed System

Model Phases

- User authentication
- System connection establishment
- System connection establishment response
- Application event entry generation
- Applications termination
- System connection termination

A Valid Authentication Trace

steena logged in at 2007-12-30 06:46:14 with user ID 3484697061757909308 with user fingerprint 94173252

C:\Program Files\Internet Explorer\IEXPLORE.exe launched at 2007-12-30 06:47:26 with ID 4384844220178160764 with fingerprint 223266966

C:\Program Files\Internet Explorer\IEXPLORE.exe terminated at 2007-12-30 06:51:33 with ID 4384844220178160764 with fingerprint 223266966
Research is Important Because
• Makes electronic evidence authentic, admissible, believable, and reliable
• Satisfies evidence authentication requirements as defined by the Computer Crime and Intellectual Property Section, Criminal Division, United States Department of Justice.
• Provides copies of evidence
• Makes every entity forensically responsible
• Serves as a means of recovery and backup

Syslog Security Research
• Secure audit log research
• Secure syslog variants—Syslog-Sign and Syslog Auth
• Architecture for forensically securing logs

Phase 1
• User authentication
• Login credentials

Phase 2
System Connection Establishment
{systemprint, userprint, randomNumber}
• randomNumber
• Randomness?

Phase 3
System Connection Establishment Response
{systemprint, userprint, randomNumber-1, NONCE}KSystem
• Nonce
• KSystem—Public key
• randomNumber-1

Phase 4
Application Event Entry Generation
{userprint, app_print}NONCE
• Application fingerprint
• User fingerprint
• NONCE
Phase 5

Application Termination

{terminate message, system_print, app_print}NONCE

- Explicit terminate message
- System fingerprint
- Application fingerprint

Phase 6

System Connection Termination

{terminate message, user_print, system_print}NONCE

- User logs off
- Explicit logging of service termination

Digital Fingerprints

- Identification
- Verification
- Equivalent to their real world counterpart

The Robert Sedgwick (RS) Algorithm

- The RS algorithm is a general-purpose hashing algorithm

\[
hash = hash^{t \cdot \langle \gg p \rangle \cdot \langle \ll q \rangle}
\]

- Least number of collisions [6]

User Fingerprints

- Username and password
- System mac address
- Login time

- An example of a user fingerprint yielded by this method ➔ 15590563

Application Fingerprints

- Launch time
- Username
- System mac address
- Application identifier

- An example of an application print yielded by this method ➔ 76274804
**System Fingerprints**

- The number of processors
- Disk space
- System mac address
- CPU ID
- Installed applications
- Disk drive identifier, serial number

An example of a system fingerprint yielded by this method → 161044579

---

**Authentication Traces**

- An umbrella activity
- Generated on every system on the network
- Records the generation fingerprints
- Serves as a forensically friendly copy of syslog entries.
- Authentication server has copies

An example:

---

**Authentication Trace Example**

(1) steena logged in at 2007-12-30 06:46:14 with user ID 3488469706175790508 with user fingerprint 88726020
The system print is 94173252

C:\Program Files\Internet Explorer\IEXPLORE.exe launched at 2007-12-30 06:47:26 with ID 4384844220178160764 with fingerprint 223266966

C:\Program Files\Internet Explorer\IEXPLORE.exe terminated at 2007-12-30 06:51:13 with ID 4384844220178160764 with fingerprint 223266966

(2) Incorrect login with username: ghost occurred at 2007-12-28 23:01:16 with userID 2221344687639655740 with userprint 238806054

---

**Forensically Viable Because**

- Protected against loss
- Evidence can be found within logs
- Electronic activity is documented for additional judicial scrutiny

---

**Attack Track Back—Syslog Entries**

The corresponding syslog entries with host name, facility, priority, message, and timestamp.

<table>
<thead>
<tr>
<th>Hostname</th>
<th>Facility</th>
<th>Priority</th>
<th>Message</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>localhost</td>
<td>4</td>
<td>10</td>
<td>steena has logged in</td>
<td>2008-02-06 12:49:33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hostname</th>
<th>Facility</th>
<th>Priority</th>
<th>Message</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>localhost</td>
<td>6</td>
<td>12</td>
<td>C:\Program Files\Internet Explorer\IEXPLORE.exe launched at</td>
<td>2008-02-06 12:49:41</td>
</tr>
</tbody>
</table>

---

**Attack Track Back—Authentication Traces**

steena logged in at 2008-02-06 12:49:33 with user ID 7524389880967786033 with user fingerprint 155990563 the system print is 161044579

C:\Program Files\Internet Explorer\IEXPLORE.exe launched at 2008-02-06 12:49:41 with ID 1524843500148472672 with fingerprint 88504721
Verifying the Fingerprints

By using:
- Authentication traces
- Syslog entries

Tracking Bad Logins

Incorrect login with username: steena occurred at 2008-02-07 04:50:02 with userID 56032638045929763 with userprint 188996098
Incorrect login with username: steena occurred at 2008-02-07 04:50:28 with userID 8936243886107892818 with userprint 188996200
Incorrect login with username: steena occurred at 2008-02-07 04:50:43 with userID 2404564924573438423 with userprint 188996163

Resilience of the Method Against—

- The challenge response mechanism
- Syslog
  - Syslog truncation
  - Denial of service
  - Abuse of privileges
- Valid application updates
- Rogue application identification

Attacking the Challenge Response Mechanism

System Connection Establishment
\{systemprint, userprint, randomNumber\}

Attack:
- Sniffing
- Replay

Prevention:
- Authentication
- Fingerprints

Attacking the Challenge Response Mechanism

Application Event Entry Generation
\{userprint, app_print, NONCE\}

Attack:
- Replay

Prevention:
NONCE—encrypted and transmitted
Attacking the Challenge Response Mechanism

**Application Termination**

{terminatemessage, app_print, systemprint}NONCE

**Attack:**

Sniffing

**Prevention:**

Encryption

**System Connection Termination**

{terminatemessage, userprint, systemprint}NONCE

**Attack:**

Man-in-the-middle/Sniffing

**Prevention:**

NONCE (previous phase)

Attacking the System Log File/Server

**Attacks:**

Truncation
Spurious entry injection
DOS

**Prevention:**

Authentication traces
Separation of syslogs and authentication traces

With Regard to Applications

**Concerns:**

Identifying Rogue Applications
Application Updates

**Prevention:**

Application IDS
Application Fingerprints
Authentication Traces

Conclusions

- Forensic techniques are in their infancy
- These are issues needing solutions, recommendations, publications
  - Submit ideas for next year
- Does the solution fit the need
  - Technological + Legal
Data Sets

• Self created
• honeynet.org
• DFRWS

Conferences for Digital Forensics

• DFRWS
  – Heavily practitioner oriented
• SADFE
  – Integrates researchers, lawyers, practitioners
• Most security conferences accept some such papers