Detecting Data Theft Using Stochastic Forensics

Jonathan Grier
DFRWS 2011
Data Exfiltration

I’ve received a number of questions both via e-mail and from customers, asking about data exfiltration. In the vast majority of cases, someone has a system (or an image acquired from a system) and wants to know what data was copied off that system, possibly onto a removable storage device. The fact of the matter is that there are a number of means by which a user can copy data off a system, such as by attaching files to Web-based e-mails, using the built-in File Transfer Protocol (FTP) client, and so forth. When you’re looking for indications or “evidence” that files were copied from the system to removable media (e.g., a thumb drive, iPod, etc.), the simple fact is that at this time, there are no apparent artifacts of this process, and you would need to acquire and analyze both pieces of media (i.e., the system that was the source, and the removable media that was the target). Artifacts of a copy operation, such as using the copy command or drag-and-drop, are not recorded in the Registry, or within the file system, as far as I and others have been able to determine.

Harlan Carvey, *Windows Forensic Analysis*, 2009
Data Exfiltration

I've received a number of questions both via e-mail and from customers, asking about data exfiltration. In the vast majority of cases, someone has a system (or an image acquired from a system) and wants to know what data was copied off that system, possibly onto a removable storage device. The fact of the matter is that there are a number of means by which a user can copy data off a system, such as by attaching files to Web-based e-mails, using the built-in File Transfer Protocol (FTP) client, and so forth. When you’re looking for indications or “evidence” that files were copied from the system to removable media (e.g., a thumb drive, iPod, etc.), the simple fact is that at this time, there are no apparent artifacts of this process, and you would need to acquire and analyze both pieces of media (i.e., the system that was the source, and the removable media that was the target). Artifacts of a copy operation, such as using the copy command or drag-and-drop, are not recorded in the Registry, or within the file system, as far as I and others have been able to determine.

Harlan Carvey, *Windows Forensic Analysis*, 2009

No Artifacts = No Forensics
Data Exfiltration

I’ve received a number of questions both via e-mail and from customers, asking about data exfiltration. In the vast majority of cases, someone has a system (or an image acquired from a system) and wants to know what data was copied off that system, possibly onto a removable storage device. The fact of the matter is that there are a number of means by which a user can copy data off a system, such as by attaching files to Web-based e-mails, using the built-in File Transfer Protocol (FTP) client, and so forth. When you’re looking for indications or “evidence” that files were copied from the system to removable media (e.g., a thumb drive, iPod, etc.), the simple fact is that at this time, there are no apparent artifacts of this process, and you would need to acquire and analyze both pieces of media (i.e., the system that was the source, and the removable media that was the target). Artifacts of a copy operation, such as using the copy command or drag-and-drop, are not recorded in the Registry, or within the file system, as far as I and others have been able to determine.

Harlan Carvey, *Windows Forensic Analysis*, 2009

No Artifacts = No Forensics???
Project Aurora
  - Engineering
    - Design
      - Blueprint.dwg
    - Tests
      - Heat.xls
      - Vibration.xls
  - Vendors
    - MCarthySmith.doc
    - TBAC-Systems.doc
Access timestamps updates during:

Routine access

1. 9:13:01 AM
2. 9:13:03 AM
3. 9:13:04 AM
4. 9:13:06 AM
5. 9:17:25 AM
6. 9:21:47 AM
Project Aurora
- Engineering
  - Design
    - Blueprint.dwg
  - Tests
    - Heat.xls
    - Vibration.xls
- Vendors
  - MccarthySmith.doc
  - TBAC-Systems.doc
Access timestamps updates during:

Copying a folder

1. 9:13:01 AM
2. 9:13:01 AM
3. 9:13:01 AM
4. 9:13:01 AM
5. 9:13:03 AM
6. 9:13:03 AM
7. 9:13:04 AM
8. 9:13:05 AM
9. 9:13:05 AM
10. 9:13:05 AM

Routine access

1. 9:13:01 AM
2. 9:13:03 AM
3. 9:13:04 AM
4. 9:13:06 AM
5. 9:17:25 AM
6. 9:21:47 AM
<table>
<thead>
<tr>
<th>Copying Folders</th>
<th>Routine Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonselective</td>
<td>Selective</td>
</tr>
<tr>
<td>All subfolders and files accessed</td>
<td></td>
</tr>
<tr>
<td>Temporally continuous</td>
<td>Temporally irregular</td>
</tr>
<tr>
<td>Recursive</td>
<td>Random order</td>
</tr>
<tr>
<td>Directory accessed before its files</td>
<td>Files can be accessed without directory</td>
</tr>
</tbody>
</table>
COPIED

NOT COPIED
### No Artifacts

**Yes Forensics**

"slap-your-head-and-say-'doh-wish-l'd-thought-of-that"

-- an anonymous reviewer

<table>
<thead>
<tr>
<th>Copying Folders</th>
<th>Routine Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective</td>
<td>Random order</td>
</tr>
<tr>
<td>Temporally irregular</td>
<td>Files can be accessed without directory</td>
</tr>
<tr>
<td>Temporarily continuous</td>
<td>Recursive</td>
</tr>
<tr>
<td>All folders and files accessed</td>
<td>Recursive</td>
</tr>
<tr>
<td>Non-selective</td>
<td>Selective</td>
</tr>
<tr>
<td>Directory accessed before its files</td>
<td>Temporarily irregular</td>
</tr>
</tbody>
</table>
Not so fast...

1. Timestamps are overwritten very quickly

2. There are other nonselective, recursive activities (besides copying)
Not so fast...

1. Timestamps are overwritten *very quickly*

   Can we use this method months later?

   On a heavily used system?

   Won’t most of the timestamps have been overwritten?
Not so fast...

1. Timestamps are overwritten *very quickly*

**YES!** Can we use this method months later?

**YES!** On a heavily used system?

*Not really!* Won’t most of the timestamps have been overwritten?
Two observations:

1. Timestamps values can increase, but never decrease.

2. A lot of files just collect dust. Most activity is on a minority of files.
The vast majority of files on two fairly typical Web servers have not been used at all in the last year. Even on an extraordinarily heavily used (and

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Over one year:</td>
<td>76.6</td>
<td>75.9</td>
<td>10.9</td>
</tr>
<tr>
<td>Six months to one year:</td>
<td>7.6</td>
<td>18.6</td>
<td>7.2</td>
</tr>
</tbody>
</table>

Farmer & Venema, *Forensic Discovery*, 2005

---

**Table 1.1 Percentage of files read or executed recently for a number of Internet servers**

**Pareto Principle**
- 80% of your profits come from 20% of...
At $t_{\text{copying}}$:

• All files have $\text{access\_timestamp} = t_{\text{copying}}$
At $t_{\text{copying}}$:
- All files have $\text{access\_timestamp} = t_{\text{copying}}$

Several weeks later:
- All files have $\text{access\_timestamp} \geq t_{\text{copying}}$
At $t_{\text{copying}}$:
• All files have $\text{access\_timestamp} = t_{\text{copying}}$

Several weeks later:
• All files have $\text{access\_timestamp} \geq t_{\text{copying}}$
• Many files still have $\text{access\_timestamp} = t_{\text{copying}}$
Histogram of access timestamps

FolderB (not copied)

After 300 days of simulated activity
Folder A (copied)

Cutoff cluster

(No residual activity visible)

Folder B (not copied)

Residual activity visible

(No cutoff cluster)
Copying creates a **cutoff cluster**

- **cutoff** – No file has timestamp \(< t_{\text{cluster}}\)
- **cluster** – Many files have timestamp \(= t_{\text{cluster}}\)
An actual investigation:

### Table 2 – Metrics applied to field investigation. All values are over range \( t_{\text{investigation}} - 180\text{days}, t_{\text{investigation}} \) unless otherwise noted.

<table>
<thead>
<tr>
<th></th>
<th>FolderQ</th>
<th>FolderR</th>
<th>FolderS</th>
<th>FolderT</th>
<th>FolderU</th>
</tr>
</thead>
<tbody>
<tr>
<td>A priori hypothesis</td>
<td>Suspected of being copied</td>
<td>Not suspected of being copied</td>
<td>≈800</td>
<td>≈300</td>
<td>≈50</td>
</tr>
<tr>
<td>(</td>
<td>D(f))</td>
<td></td>
<td>≈6000</td>
<td>≈7000</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Cluster (_t)</td>
<td>&gt;0.3 (at (t_1))</td>
<td>&gt;0.9 (at (t_2))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indication</td>
<td>Copied at (t_1)</td>
<td>Copied at (t_2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{Mag}_t)</td>
<td>&gt;5000 ((t = t_1))</td>
<td>&gt;6000 ((t = t_2))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(</td>
<td>\text{Abn}_t</td>
<td>)</td>
<td>&gt;50000 ((t = t_1))</td>
<td>&gt;20000 ((t = t_2))</td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>Suspicion supported</td>
<td>Subsequent investigation determined this copying was authorized</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>forensically</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Jonathan Grier, *Detecting Data Theft Using Stochastic Forensics*, DFRWS 2011
Leading researchers have called to move from: “What data can we find?”

To:

“What did this person do?”
Classical Forensics:

- Look at the Surviving Data
- Reconstruct Previous Data
- This previous data is our deliverable.
Classical Forensics:

Look at the Surviving Data → Reconstruct Previous Data → This previous data is our deliverable.

Stochastic Forensics:

What do I want to know about? → What behavior is associated? → How does that behavior affect the system? → Measure those effects. Draw a (quantifiable) inference.
Aren’t there other recursive access patterns besides copying?

*Affirming the consequent*
A → B doesn’t prove B → A.

The *absence* of a cutoff cluster can disprove copying, but the *existence* can’t prove copying.

Perhaps they ran `grep`. 
Indeed, there are!

Affirming the consequent
A → B doesn’t prove B → A.

The absence of a cutoff cluster can disprove copying, but the existence can’t prove copying.

Who’s trying to prove anything?
Investigate! One clue leads to another until the case unravels.

Abductive reasoning
An unusual observation supports inferring a likely cause.

Indeed!
Check if grep is installed, if they’ve ever run it before, or after, on any folder.
Check why they were still in the building at 11 PM.

Perhaps they ran grep.
Forensics

WHY PROGRAMMING IS A GOOD MEDIUM FOR EXPRESSING POORLY UNDERSTOOD AND SLOPPILY-FORMULATED IDEAS.

-- Marvin Minsky, MIT, 1967
Our general philosophy recommends greater understanding instead of higher levels of certainty, which could potentially make such methodology more suspect in a court of law. Paradoxically, however, the uncertainty—primarily in the data collection methods—can actually give a greater breadth of knowledge and more confidence in any conclusions.

Farmer & Venema, *Forensic Discovery*, 2005
Open Questions
(i.e. a request for help)

1. Scientific testing

2. Probability value

3. Fingerprinting
   We can distinguish copying from grep!

4. What other questions can stochastic forensics address?
   Let’s find sloppy questions and answer them less precisely!
I’m very interested in hearing feedback, ideas, and questions.

Please share them with me here at DFRWS.

Or, if we miss each other:

Jonathan Grier
443.501.4044 x1
jgrier at vesaria.com