Employee Data Theft Case Study

Jonathan Grier
ACSAC 2011
Concerning Confidentiality

To preserve client confidentiality, this case’s circumstantial information (names, places, dates, and settings) has been omitted or altered.

The data and techniques presented have not been altered.
Can you find the data thief?
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???
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
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
<table>
<thead>
<tr>
<th>Copying Folders</th>
<th>Routine Access</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nonselective</strong></td>
<td><strong>Selective</strong></td>
</tr>
<tr>
<td>All subfolders and files accessed</td>
<td></td>
</tr>
<tr>
<td><strong>Temporally continuous</strong></td>
<td><strong>Temporally irregular</strong></td>
</tr>
<tr>
<td><strong>Recursive</strong></td>
<td><strong>Random order</strong></td>
</tr>
<tr>
<td>Directory accessed before its files</td>
<td>Files can be accessed without directory</td>
</tr>
</tbody>
</table>
No Artifacts
Yes Forensics

“slap-your-head-and-say-'doh-wish-I'd-thought-of-that”

-- an anonymous colleague
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

\textcolor{red}{YES!} Can we use this method months later?

\textcolor{red}{YES!} On a heavily used system?

\textcolor{green}{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 1.1 Percentage of files read or executed recently for a number of Internet servers

<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
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
Data from investigation:

<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>n̄ = 6000</td>
<td>≥ 7000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Cluster</td>
<td>&gt;0.3 (at t = t₁)</td>
<td>&gt;0.9 (at t = t₂)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indication</td>
<td>Copied at t₁</td>
<td>Copied at t₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magₜ</td>
<td>&gt;5000 (t = t₁)</td>
<td>&gt;6000 (t = t₂)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;50000 (t = t₁)</td>
<td>&gt;20000 (t = t₂)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td>Suspicion supported</td>
<td>Subsequent investigation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>forensically</td>
<td>determined</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copying creates a cutoff cluster

**cutoff** – No file has timestamp < $t_{\text{cluster}}$

**cluster** – Many files have timestamp = $t_{\text{cluster}}$
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.

Perhaps they ran `grep`.

**Abductive reasoning**

An unusual observation supports inferring a likely cause.

Who’s trying to *prove* anything?

Investigate! One clue leads to another until the case unravels.

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.
Implications for the field of forensics...
Classical Forensics:

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

Look at the Surviving Data $\rightarrow$ Reconstruct Previous Data $\rightarrow$ This previous data is our deliverable.

Stochastic Forensics:

What do I want to know about? $\rightarrow$ What behavior is associated? $\rightarrow$ How does that behavior affect the system? $\rightarrow$ Measure those effects. Draw a (quantifiable) inference.
What do I want to know about?

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

What data can we find?

Classical Forensics:

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.

What did this person do?
Lesson Learned:

Forensics doesn’t really matter...

Col. John Boyd
Military Strategist
Author, *Patterns of Conflict*
For more information:

• Read my paper
  *Detecting Data Theft
  *Using Stochastic Forensics*
  

• These slides will be available at
  
  http://www.grierforensics.com/datatheft/Employee_Data_Theft_Case_Study_ACSAC.pdf

• Ask me!
  
  See next slide for my contact info
I’m very interested in hearing your feedback, ideas, and questions.

Please share them with me here at ACSAC.

Or, if we miss each other:
Jonathan Grier
443.501.4044 x1
jdgrier at grierforensics.com

PS If you’re a researcher and interested in collaborating to take this further, I’d love to speak with you!