DFRWS
DIGITAL FORENSIC RESEARCH CONFERENCE

Detecting Data Theft Using Stochastic Forensics

By

Jonathan Grier

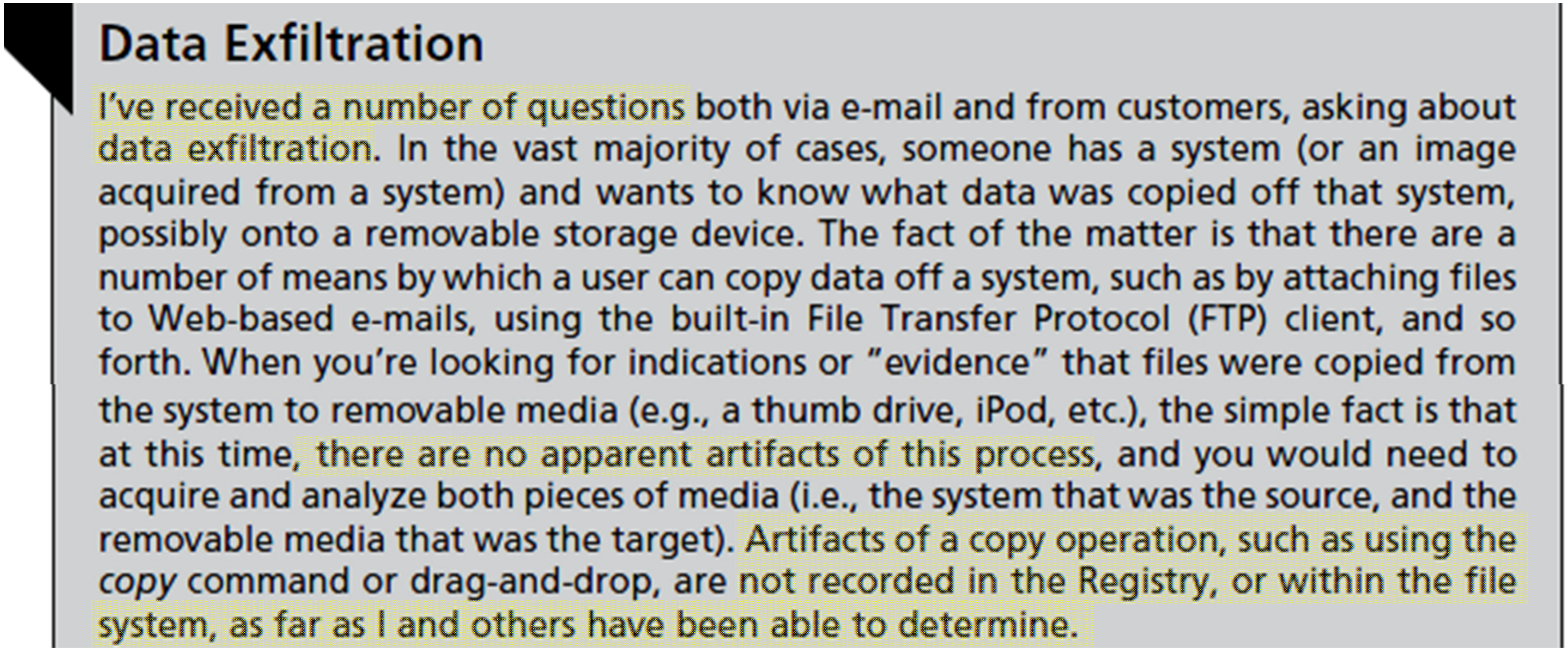
Presented At

The Digital Forensic Research Conference

DFRWS 2011 USA New Orleans, LA (Aug 1st - 3rd)

Detecting Data Theft Using Stochastic Forensics

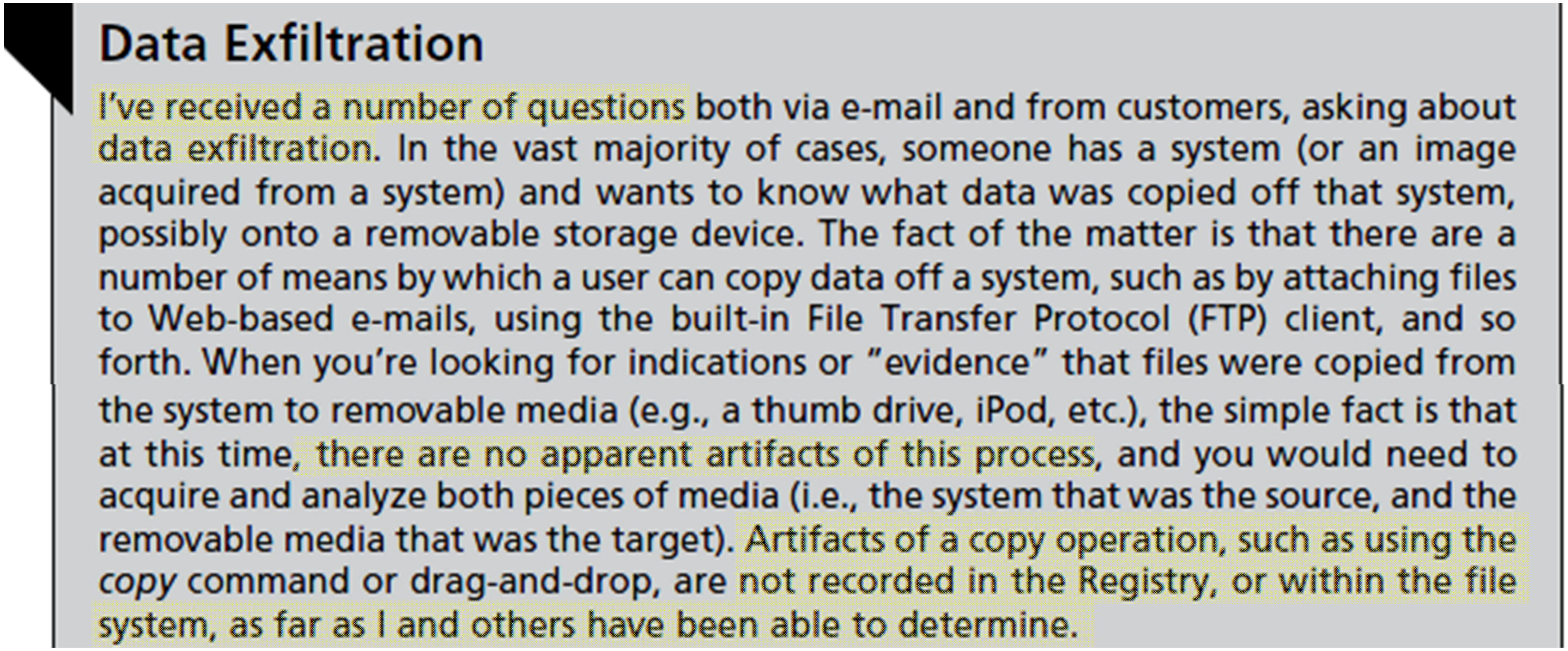
Jonathan Grier
DFRWS 2011



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

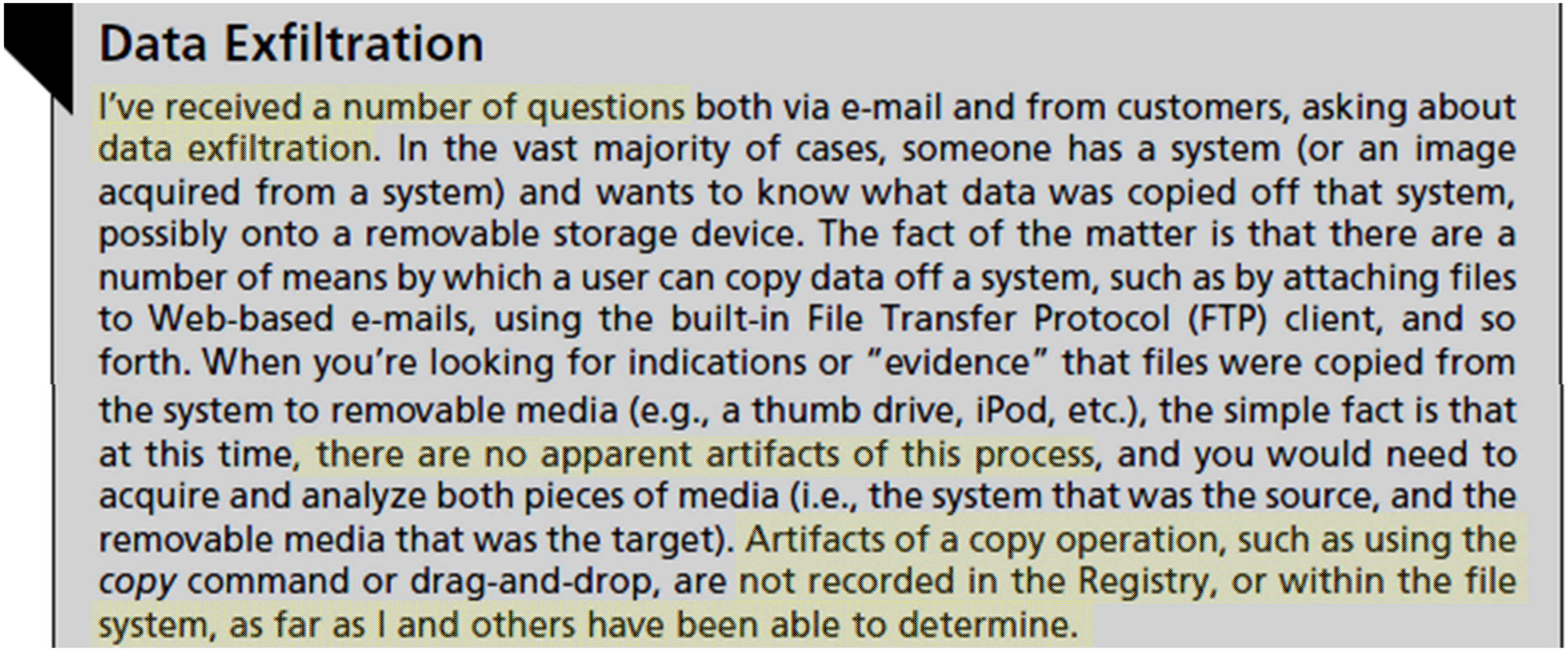


Data Exfiltration

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

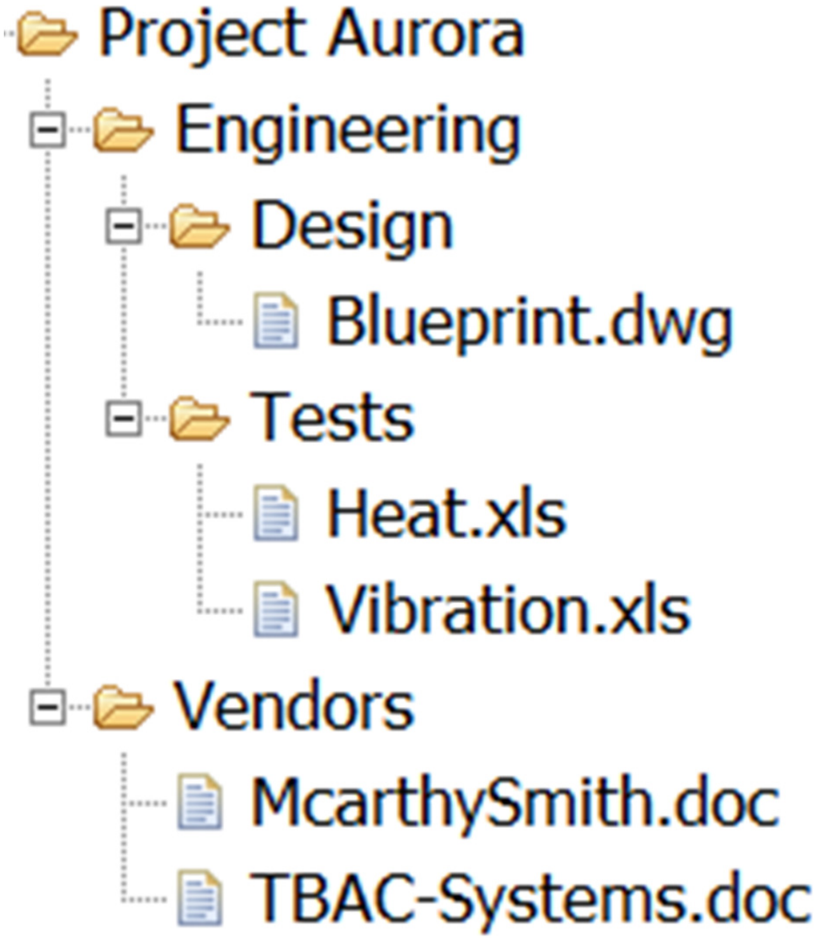


Data Exfiltration

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

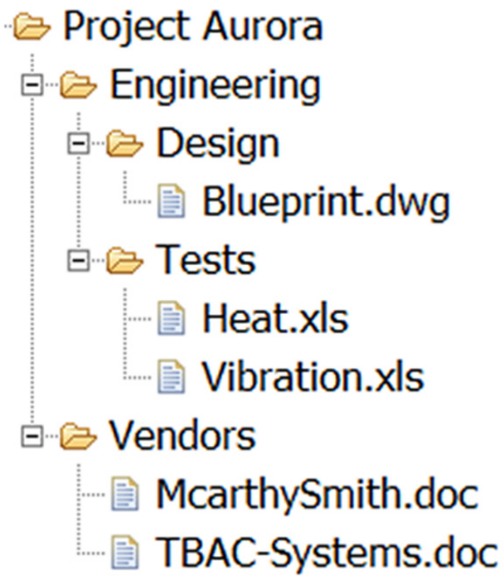
No Artifacts = No Forensics???



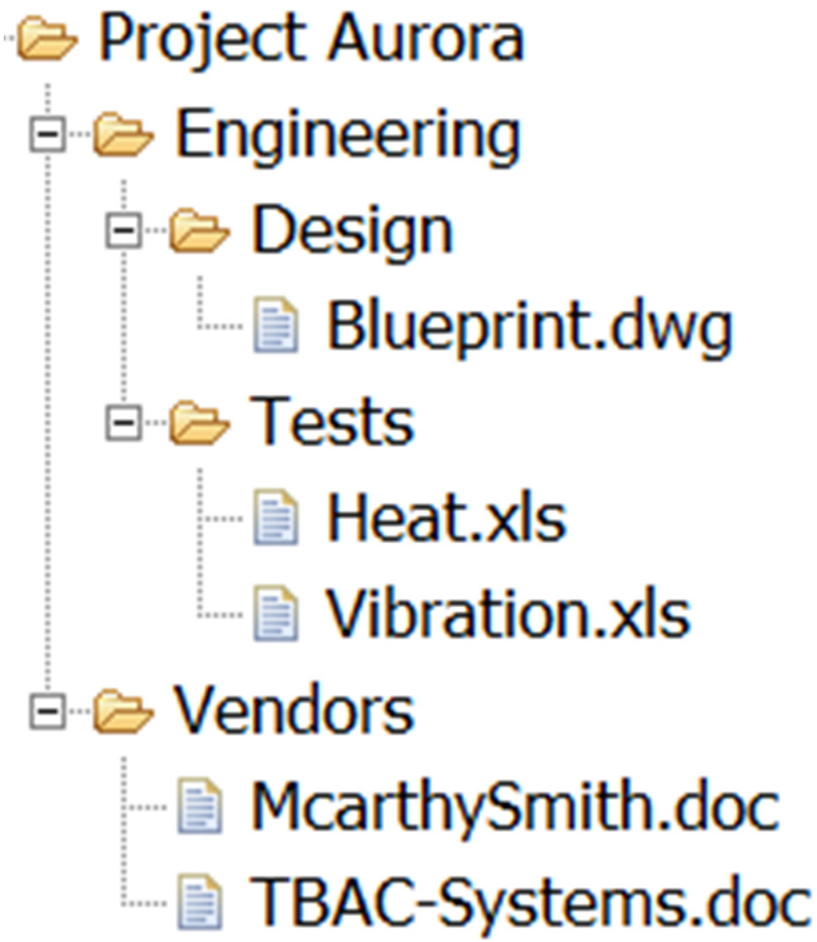
```
Project Aurora
    Engineering
        Design
            Blueprint.dwg
        Tests
            Heat.xls
            Vibration.xls
    Vendors
        McarthySmith.doc
        TBAC-Systems.doc
```

Access timestamps updates during:

Routine access



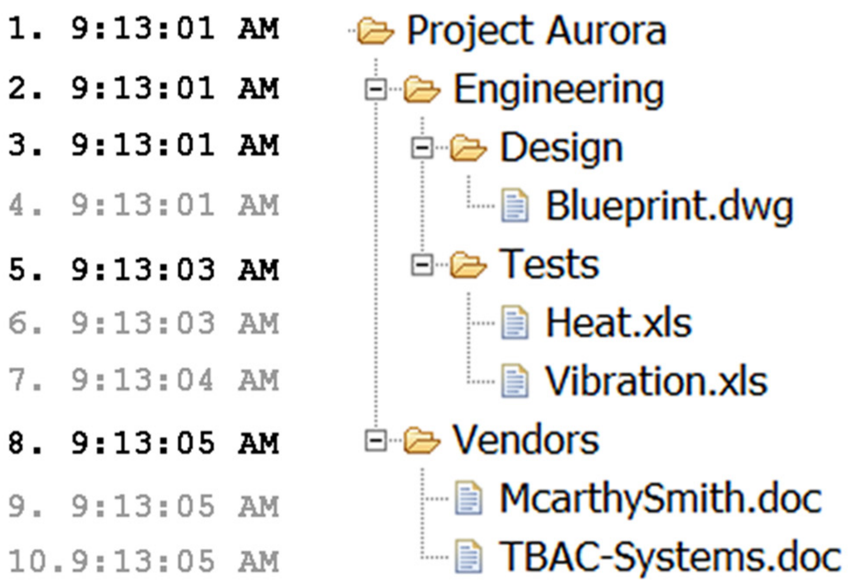
```
Project Aurora                 1. 9:13:01 AM
  Engineering                  2. 9:13:03 AM
    Design
      Blueprint.dwg            6. 9:21:47 AM
    Tests                      3. 9:13:04 AM
      Heat.xls
      Vibration.xls            4. 9:13:06 AM
  Vendors
    McarthySmith.doc           5. 9:17:25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



```
Project Aurora
  Engineering
    Design
      Blueprint.dwg
    Tests
      Heat.xls
      Vibration.xls
  Vendors
    McarthySmith.doc
    TBAC-Systems.doc
```

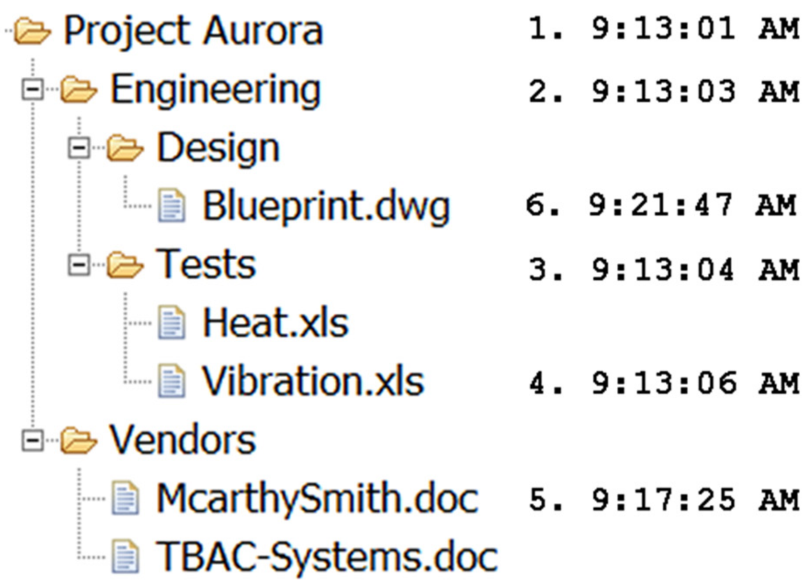
Access timestamps updates during:

Copying a folder



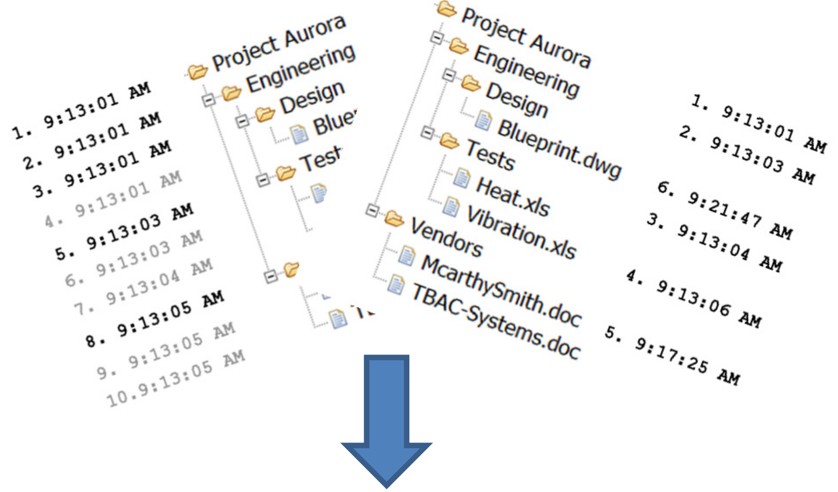
1. 9:13:01 AM	Project Aurora
2. 9:13:01 AM	Engineering
3. 9:13:01 AM	Design
4. 9:13:01 AM	Blueprint.dwg
5. 9:13:03 AM	Tests
6. 9:13:03 AM	Heat.xls
7. 9:13:04 AM	Vibration.xls
8. 9:13:05 AM	Vendors
9. 9:13:05 AM	McarthySmith.doc
10.9:13:05 AM	TBAC-Systems.doc

Routine access



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

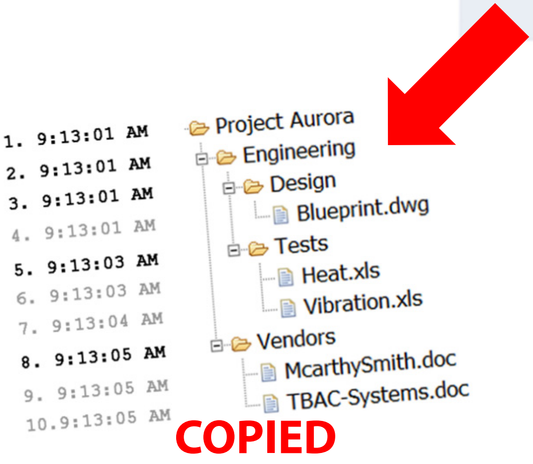
Copying Folders	Routine Access
Nonselective All subfolders and files accessed	Selective
Temporally continuous	Temporally irregular
Recursive	Random order
Directory accessed before its files	Files can be accessed without directory



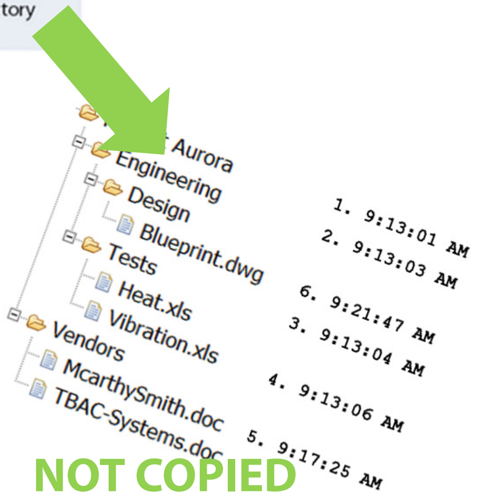
```
1. 9:13:01 AM
2. 9:13:01 AM
3. 9:13:01 AM
4. 9:13:01 AM
5. 9:13:03 AM
6. 9:13:03 AM
7. 9:13:04 AM
8. 9:13:05 AM
9. 9:13:05 AM
10.9:13:05 AM
```

```
1. 9:13:01 AM
2. 9:13:03 AM
6. 9:21:47 AM
3. 9:13:04 AM
4. 9:13:06 AM
5. 9:17:25 AM
```

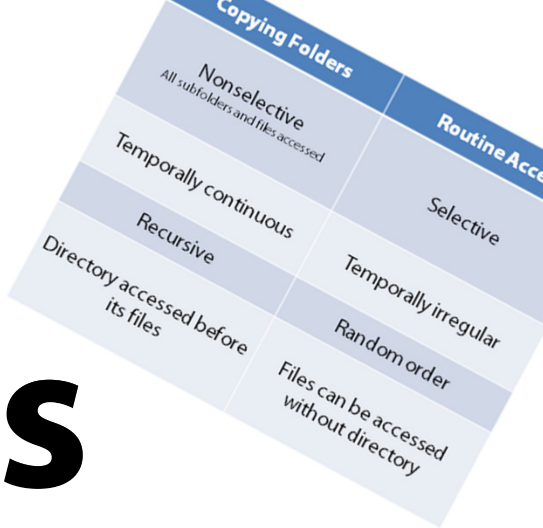
Copying Folders	Routine Access
Nonselective All subfolders and files accessed	Selective
Temporally continuous	Temporally irregular
Recursive	Random order
Directory accessed before its files	Files can be accessed without directory



```
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
```



```
1. 9:13:01 AM
2. 9:13:03 AM
6. 9:21:47 AM
3. 9:13:04 AM
4. 9:13:06 AM
5. 9:17:25 AM
```



Copying Folders	Routine Acce
Nonselective All subfolders and files accessed	Selective
Temporally continuous	Temporally irregular
Recursive	Random order
Directory accessed before its files	Files can be accessed without directory

No Artifacts

Yes Forensics

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

-- an anonymous reviewer

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?**

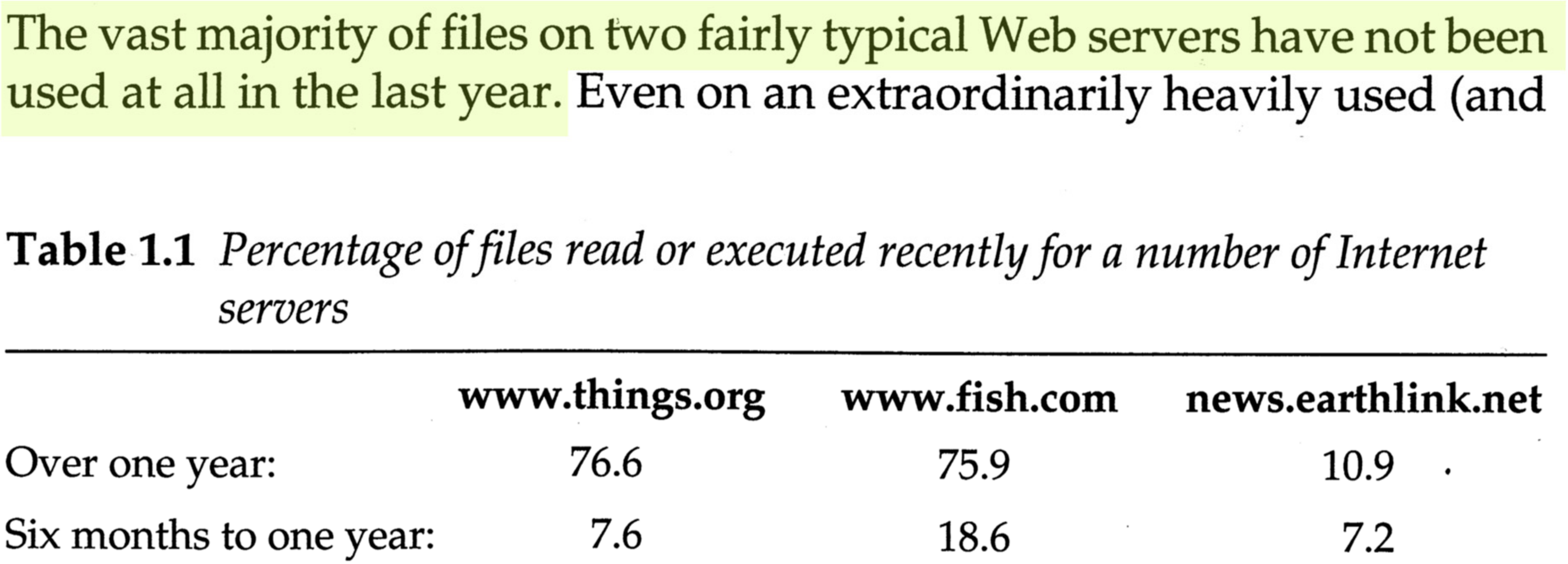
YES! **On a heavily used system?**

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.

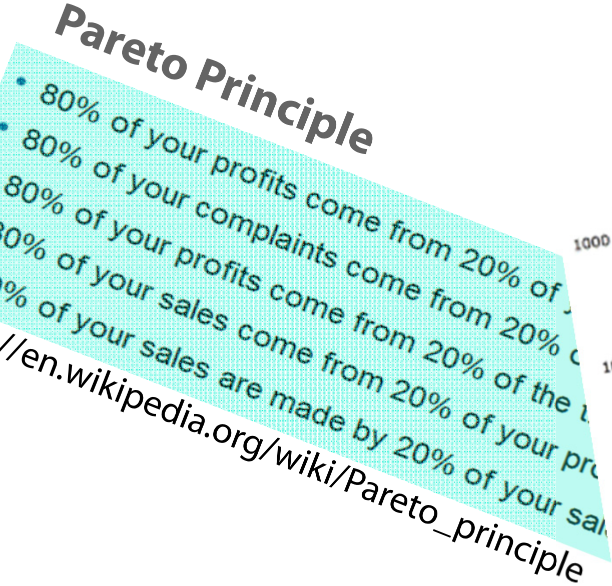


The vast majority of files on two fairly typical Web servers have not been used at all in the last year. Even on an extraordinarily heavily used (and

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

	www.things.org	www.fish.com	news.earthlink.net
Over one year:	76.6	75.9	10.9
Six months to one year:	7.6	18.6	7.2

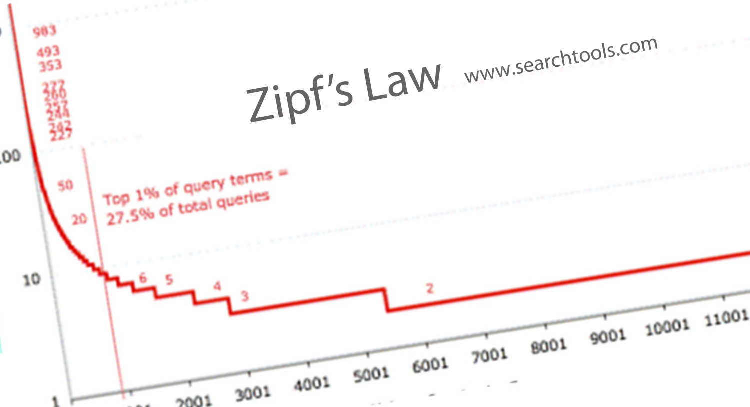
Farmer & Venema, *Forensic Discovery,* 2005



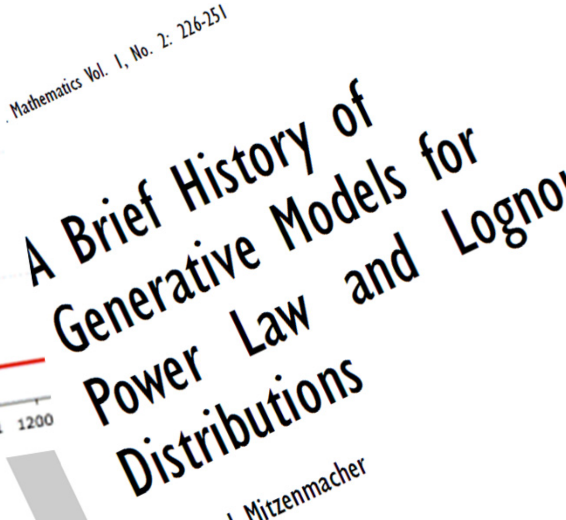
Pareto Principle

- 80% of your profits come from 20% of
- 80% of your complaints come from 20% c
- 80% of your profits come from 20% of the t
- 0% of your sales come from 20% of your pr
- % of your sales are made by 20% of your sa

//en.wikipedia.org/wiki/Pareto_principle



Zipf's Law www.searchtools.com



Mathematics Vol. 1, No. 2: 226-251

A Brief History of Generative Models for Power Law and Logno Distributions

Mitzenmacher

At $t_{copying}$:

- All files have access_timestamp = $t_{copying}$

At $t_{copying}$:

- All files have access_timestamp = $t_{copying}$

Several weeks later:

- All files have access_timestamp ≥ $t_{copying}$

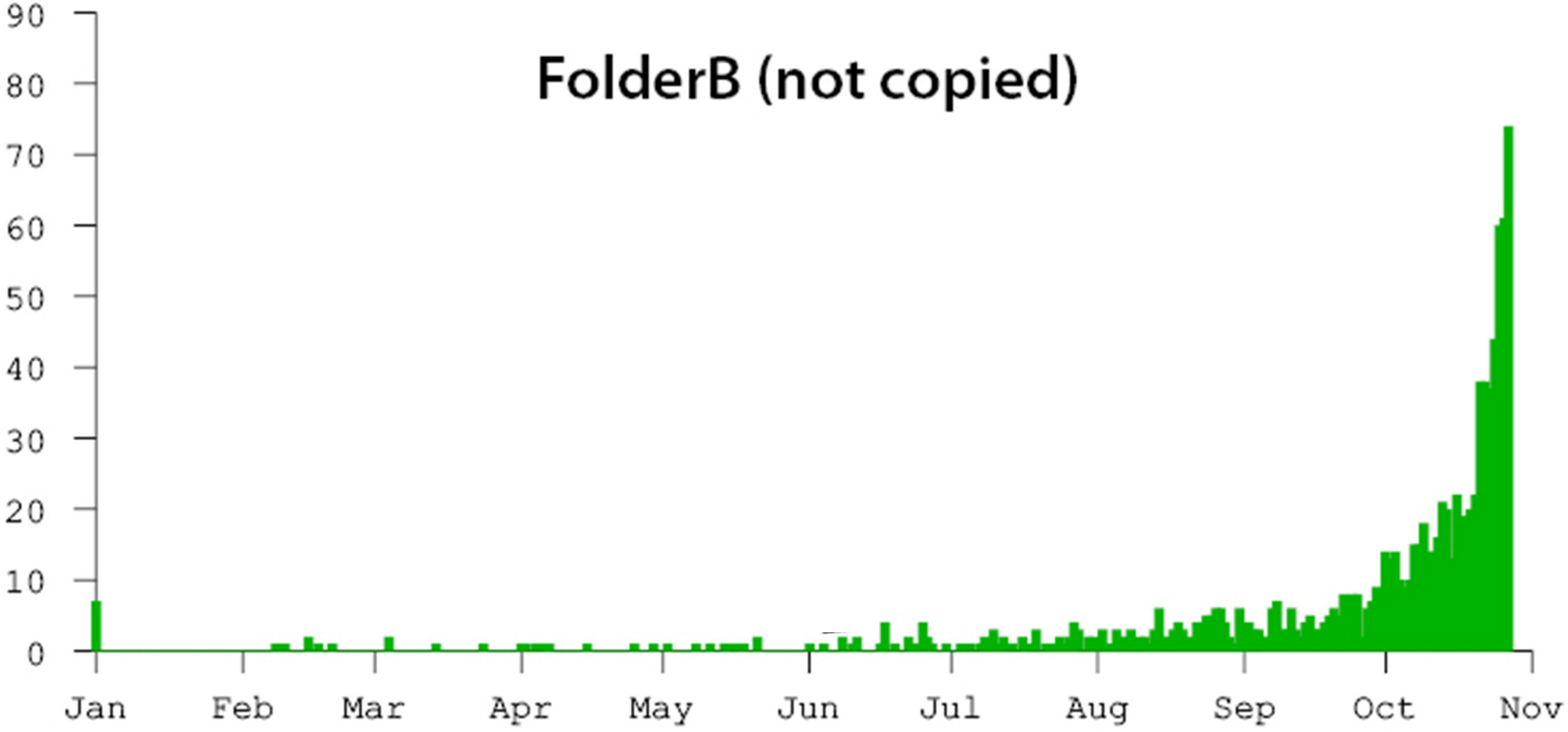
At $t_{copying}$:

- All files have access_timestamp = $t_{copying}$

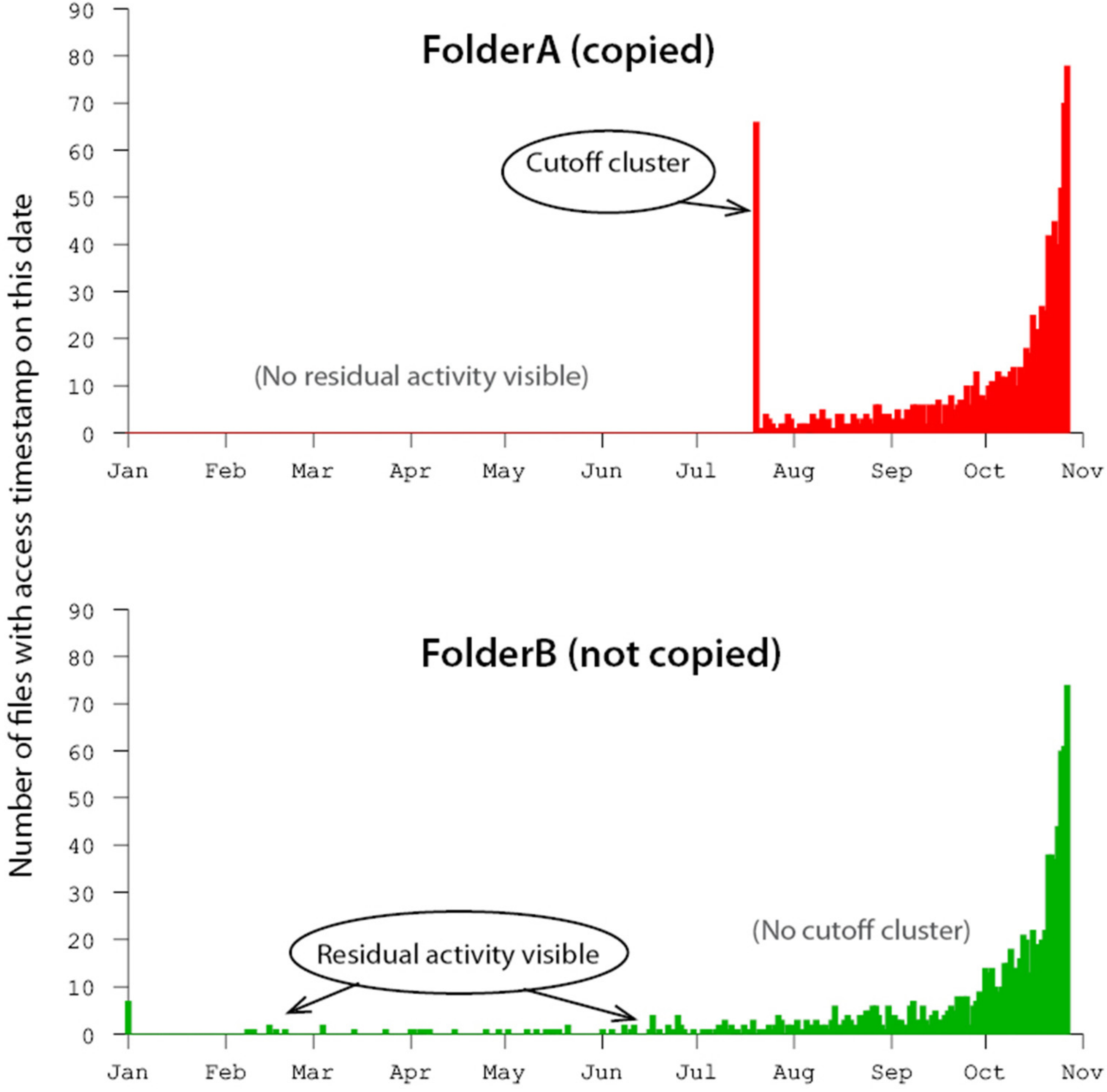
Several weeks later:

- All files have access_timestamp ≥ $t_{copying}$
- Many files still have access_timestamp = $t_{copying}$

Histogram of access timestamps



After 300 days of simulated activity


Number of files with access timestamp on this date
FolderA (copied)
Cutoff cluster
(No residual activity visible)
90
80
70
60
50
40
30
20
10
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
FolderB (not copied)
Residual activity visible
(No cutoff cluster)
90
80
70
60
50
40
30
20
10
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov

Copying creates a

cutoff cluster

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

cluster – Many files have timestamp = $t_{cluster}$

An actual investigation:

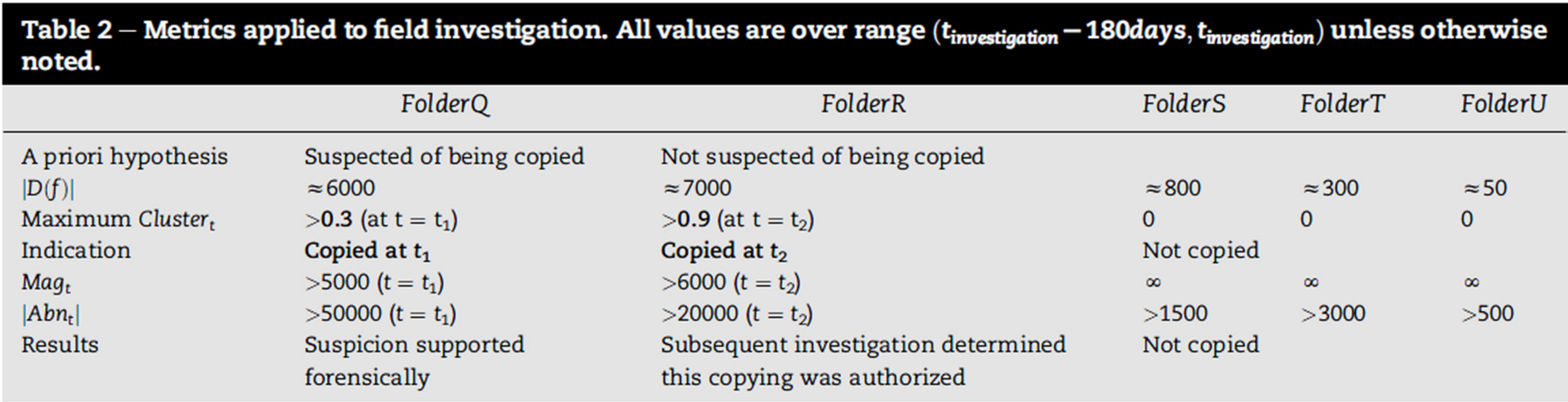
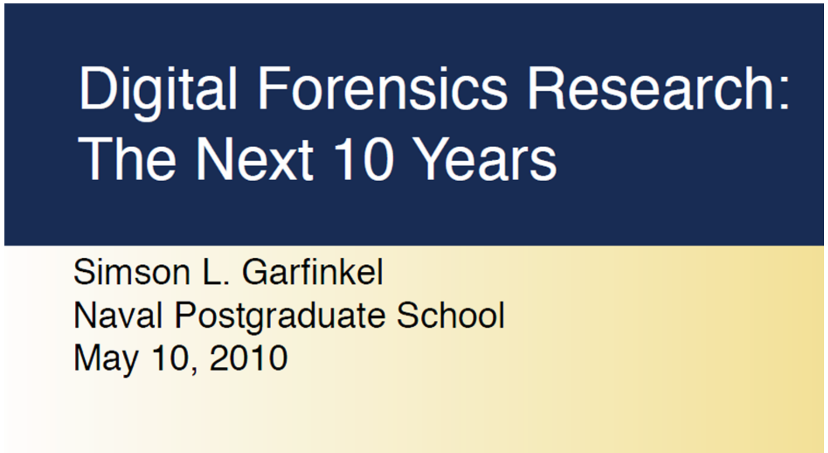


Table 2 – Metrics applied to field investigation. All values are over range ($t_{investigation} - 180days, t_{investigation}$) unless otherwise noted.

	FolderQ	*FolderR*	*FolderS*	*FolderT*	*FolderU*
A priori hypothesis	Suspected of being copied	Not suspected of being copied			
$\|D(f)\|$	≈6000	≈7000	≈800	≈300	≈50
Maximum $Cluster_t$	>**0.3** (at $t = t_1$)	>**0.9** (at $t = t_2$)	0	0	0
Indication	**Copied at t_1**	**Copied at t_2**	Not copied		
Mag_t	>5000 ($t = t_1$)	>6000 ($t = t_2$)	∞	∞	∞
$\|Abn_t\|$	>50000 ($t = t_1$)	>20000 ($t = t_2$)	>1500	>3000	>500
Results	Suspicion supported forensically	Subsequent investigation determined this copying was authorized	Not copied		

Jonathan Grier, *Detecting Data Theft Using Stochastic Forensics*, DFRWS 2011



Digital Forensics Research: The Next 10 Years

Simson L. Garfinkel
Naval Postgraduate School
May 10, 2010

Digital Forensics Research: The Good, the Bad, and the Unaddressed

by Nicole L. Beebe, Ph.D.
5th Annual IFIP WG 11.9
January 27, 2009

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.

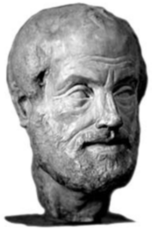
Classical Forensics:

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?



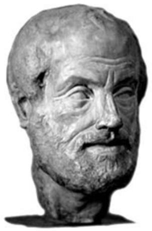
Affirming the consequent

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.

vs.

Abductive reasoning

An unusual observation supports inferring a likely cause.

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

Who's trying to *prove* anything?

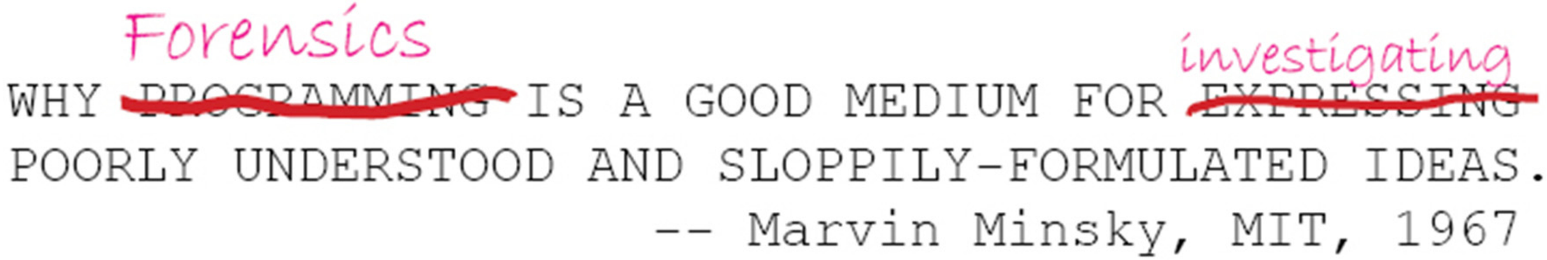
Investigate! One clue leads to another until the case unravels.

Perhaps they ran `grep`.

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.

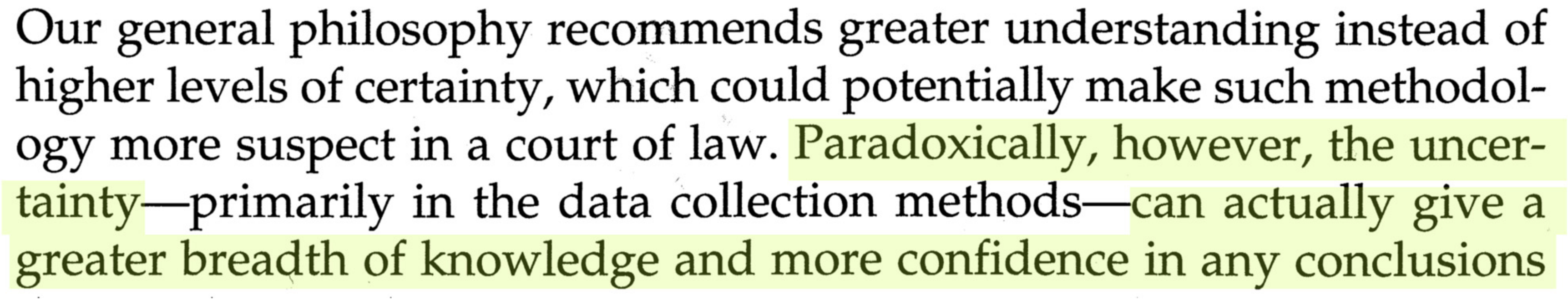
WHY ~~PROGRAMMING~~ Forensics IS A GOOD MEDIUM FOR ~~EXPRESSING~~ investigating POORLY UNDERSTOOD AND SLOPPILY-FORMULATED IDEAS.

-- Marvin Minsky, MIT, 1967



WHY ~~PROGRAMMING~~ Forensics IS A GOOD MEDIUM FOR ~~EXPRESSING~~ investigating 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