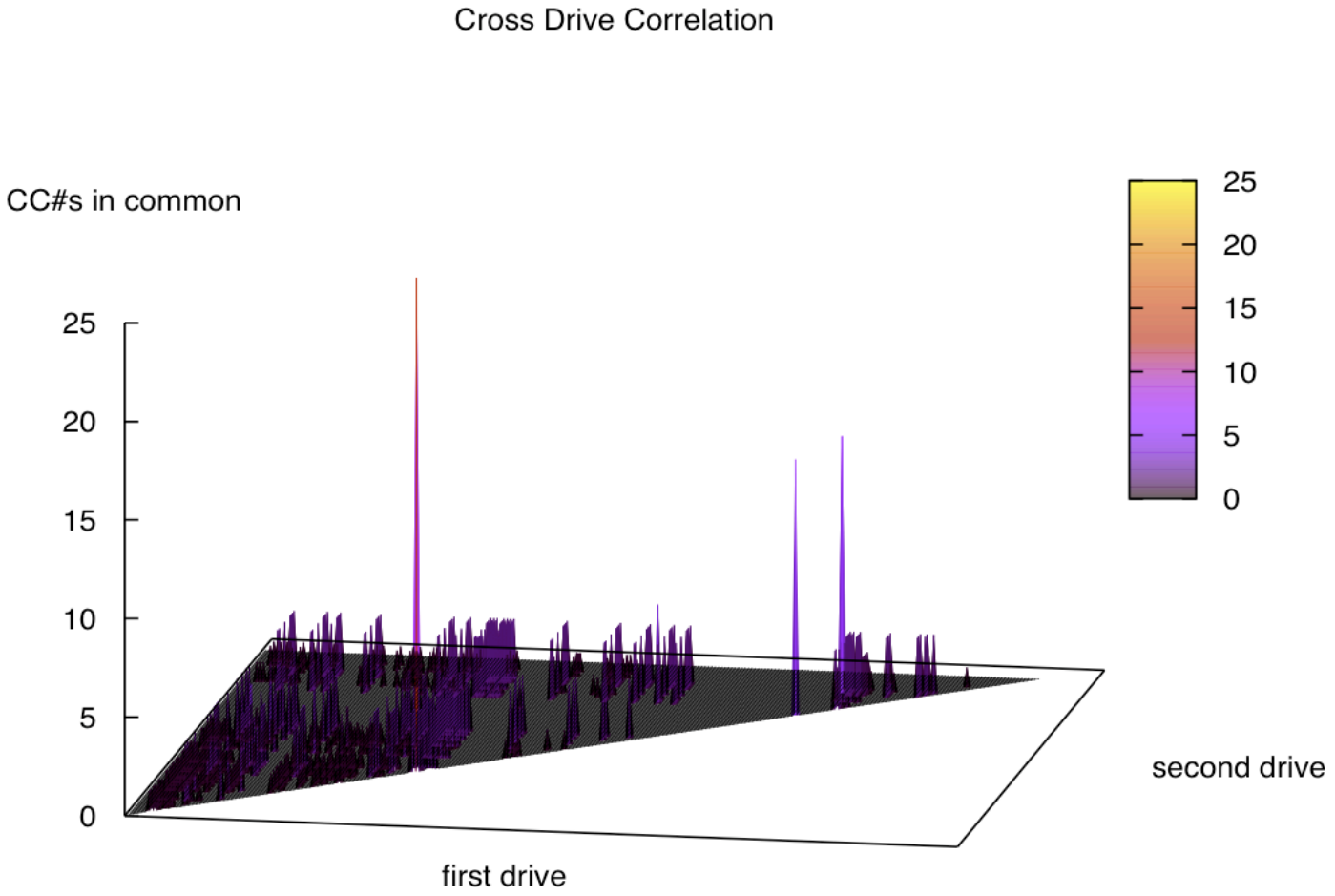
Forensic Feature Extraction and Cross-Drive Analysis



Simson L. Garfinkel

Center for Research on Computation and Society Harvard University

1:15pm, Tuesday, August 15, 2006

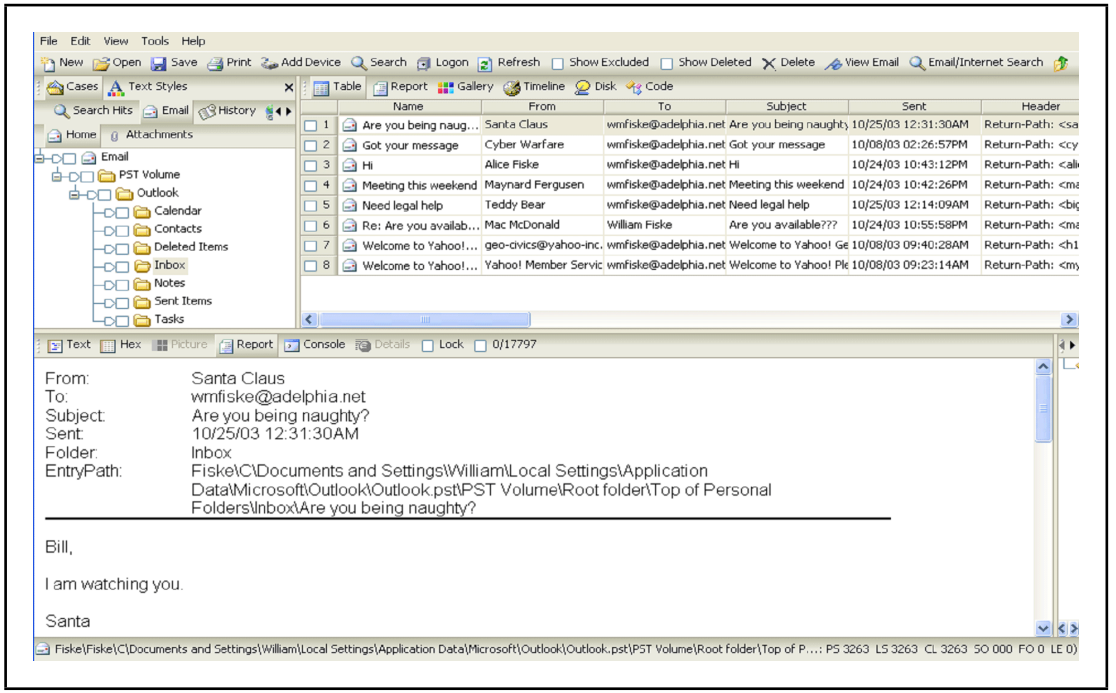
Today's forensic tools are designed for one drive at a time.

Primary Goals: Search and Recovery.

Interactive user interface.

Usage scenarios:

- Recovery of "deleted" files.
- Child porn scanning.
- Trial preparation.



Today's tools choke when confronted with hundreds or thousands of drives.

Which drives were used by my target?

Do any drives belong to the target's associates?

Who is talking to who?

Where should I start?

Police departments and intelligence agencies have thousands of drives...

Additional problems with today's tools

- Improper prioritization

  Letting priority be determined by the statute of limitations.

- Lost opportunities for data correlation

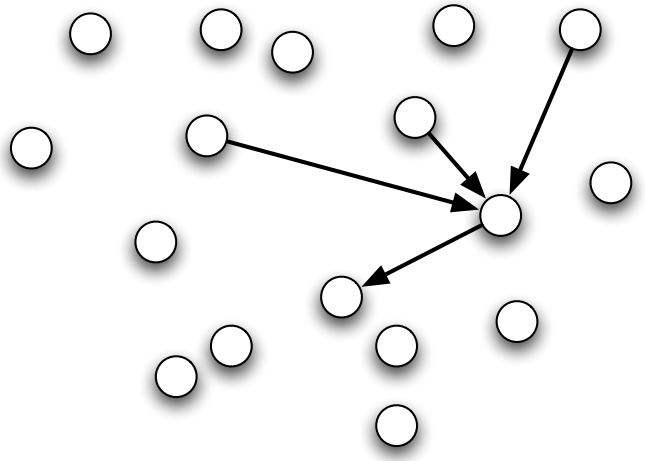
  Was a message on hard drive X sent to hard drive Y?

- Emphasis on document recovery rather than in furthering the investigation.

Correlating data *between* drives is an untapped opportunity.

How large is my target's reach?

Who is in the organization?

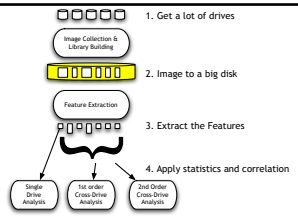


Captured drives are an ideal social network analysis.

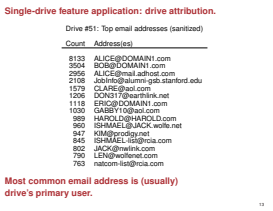
This talk introduces Cross Drive Analysis

Large scale forensics problem

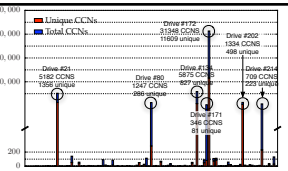
Architecture



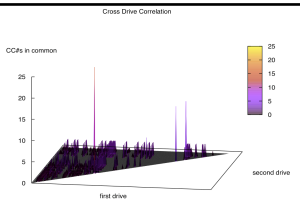
Feature extraction



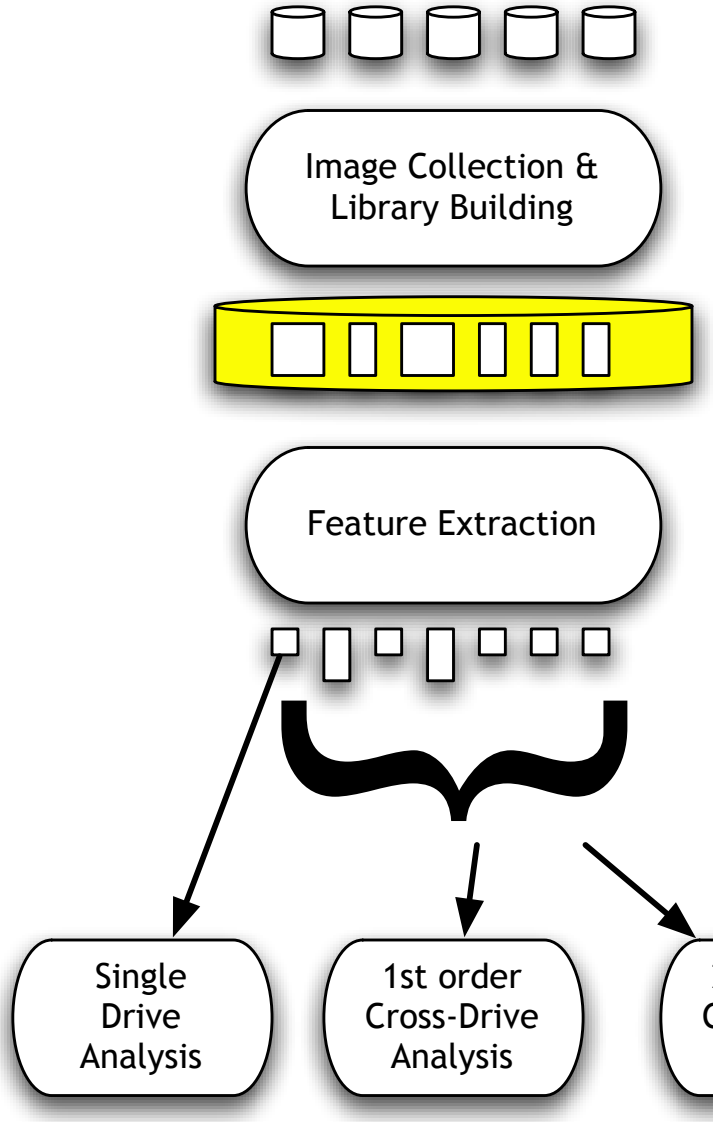
First order analysis



Second order analysis



Forensic Feature Extraction and Cross-Drive Analysis



1. Get a lot of drives
2. Image to a big disk
3. Extract the Features
4. Apply statistics and correlation

Uses of Cross-Drive Analysis

1. Automatic identification of hot drives
2. Improvements to single-drive systems
3. Identification of social network membership
4. Unsupervised social network discovery

Related Work:

- Garfinkel & Shelat, 158 drives, 2002
- FTK 2.0 — indexing multiple drives
- IntelliDact and Workshare Protect scan for confidential information

Feature extractors find *pseudo-unique* features

Pseudo-Unique characteristics:

- Long enough so collisions by chance are unlikely.
- Recognizable with regular expressions.
- Persistent over time.
- Correlated with specific documents, people or organizations.

Typical Features:

- email addresses
- Message-IDs
- Subject: lines
- Cookies
- US Social Security Numbers
- Credit card numbers
- Hash codes of drive sectors

Example: The Credit Card Number Detector.

The CCN detector scans bulk data for ASCII patterns that look like credit card numbers.

- CCNs are found in certain typographical patterns.
  (e.g. XXXX-XXXX-XXXX-XXXX
  or XXXX XXXX XXXX XXXX
  or XXXXXXXXXXXXXXXX)
- CCNs are issued with well-known prefixes.
- CCNs follow the Credit Card Validation algorithm.
- Certain numeric patterns are unlikely.
  (e.g. 4454-4766-7667-6672)

CCN detector: written in flex and C++

Scan of Drive #105: (642MB)

Test	# pass
typographic pattern	3857
known prefixes	90
CCV1	43
numeric histogram	38

Sample output:

```
'CHASE NA|5422-4128-3008-3685|    pos=13152133
'DISCOVER|6011-0052-8056-4504|    pos=13152440
.'GE CARD|4055-9000-0378-1959|    pos=13152589
BANK ONE |4332-2213-0038-0832|    pos=13152740
.'NORWEST|4829-0000-4102-9233|    pos=13153182
'SNB CARD|5419-7213-0101-3624|    pos=13153332
```

Even with the tests, there are occasional false positives.

CCN scan of Drive #115: (772MB)

Test	# pass
pattern	9196
known prefixes	898
CCV1	29
patterns	27
histogram	13

```
.................@:|444444866661O8|:<@<74444:@@@<<44   pos=82473275
............#"&'&&'|445447667667667|..050014&'4"1"&'.  pos=86493675
......221267241667&|454676676654450|&566746566726322.  pos=86507818
3..30210212676677..|30232676630232|.1.........001.01   pos=86516059
"&#&&'&41&&'645445&|454454672676632|.3............0..  pos=86523223
..........".#""#"&'|445467667227023|..............366  pos=87540819
D#9?.32400.,,+14%?B|499745255278101|*02)46+;<17756669  pos=118912826
.GGJJB...>.JJGG...G|3534554333511116|...............6  pos=197711868
%.....}}}}}}.......|444443222333345|.....}}}}}}......  pos=228610295
%6"!) .&*%,,%-0)07.|373484553420378|<67<038+.5(+0+.3.  pos=638491849
%6"!) .&*%,,%-0)07.|373484553420378|<67<038+.5(+0+.3.  pos=645913801
```

CDA Prototype System

1000 drives purchased on secondary market (1998–2006)

750 images

1.5TB data compressed.

Many different organizations.

Single-drive feature application: drive attribution.

Drive #51: Top email addresses (sanitized)

Address(es)	Count
ALICE@DOMAIN1.com	8133
BOB@DOMAIN1.com	3504
ALICE@mail.adhost.com	2956
JobInfo@alumni-gsb.stanford.edu	2108
CLARE@aol.com	1579
DON317@earthlink.net	1206
ERIC@DOMAIN1.com	1118
GABBY10@aol.com	1030
HAROLD@HAROLD.com	989
ISHMAEL@JACK.wolfe.net	960
KIM@prodigy.net	947
ISHMAEL-list@rcia.com	845
JACK@nwlink.com	802
LEN@wolfenet.com	790
natcom-list@rcia.com	763

Most common email address is (usually) drive's primary user.

Attribution histogram works even with lightly-used drives.

Extracted Email Addresses	Count on Drive #80	Total drives with address
premium-server@thawte.com	117	278
server-certs@thawte.com	104	278
CPS-requests@verisign.com	61	286
personal-premium@thawte.com	44	253
personal-basic@thawte.com	42	250
personal-freemail@thawte.com	40	250
info@netscape.com	36	58
ANGIE@ALPHA.com	32	1
BARRY@BETA.com	23	1
CHARLES@GAMMA.com	21	1
DAVE.HALL@DELTA.com	21	1
DAPHNE@UNIFORM.com	20	1
ELLY@LIMA.com	18	1
FRANK@ECHO.com	16	1
HUGH@LIMA.com	16	1
IGGY@LIMA.com	16	1
GRETTA@XYZZY.com	15	1
VISTA@SNARF.com	15	1

Email addresses found on $\approx> 20$ drives are not pseudo-unique

First Order Cross-Drive Analysis: $O(n)$ operations on feature files

Applications:

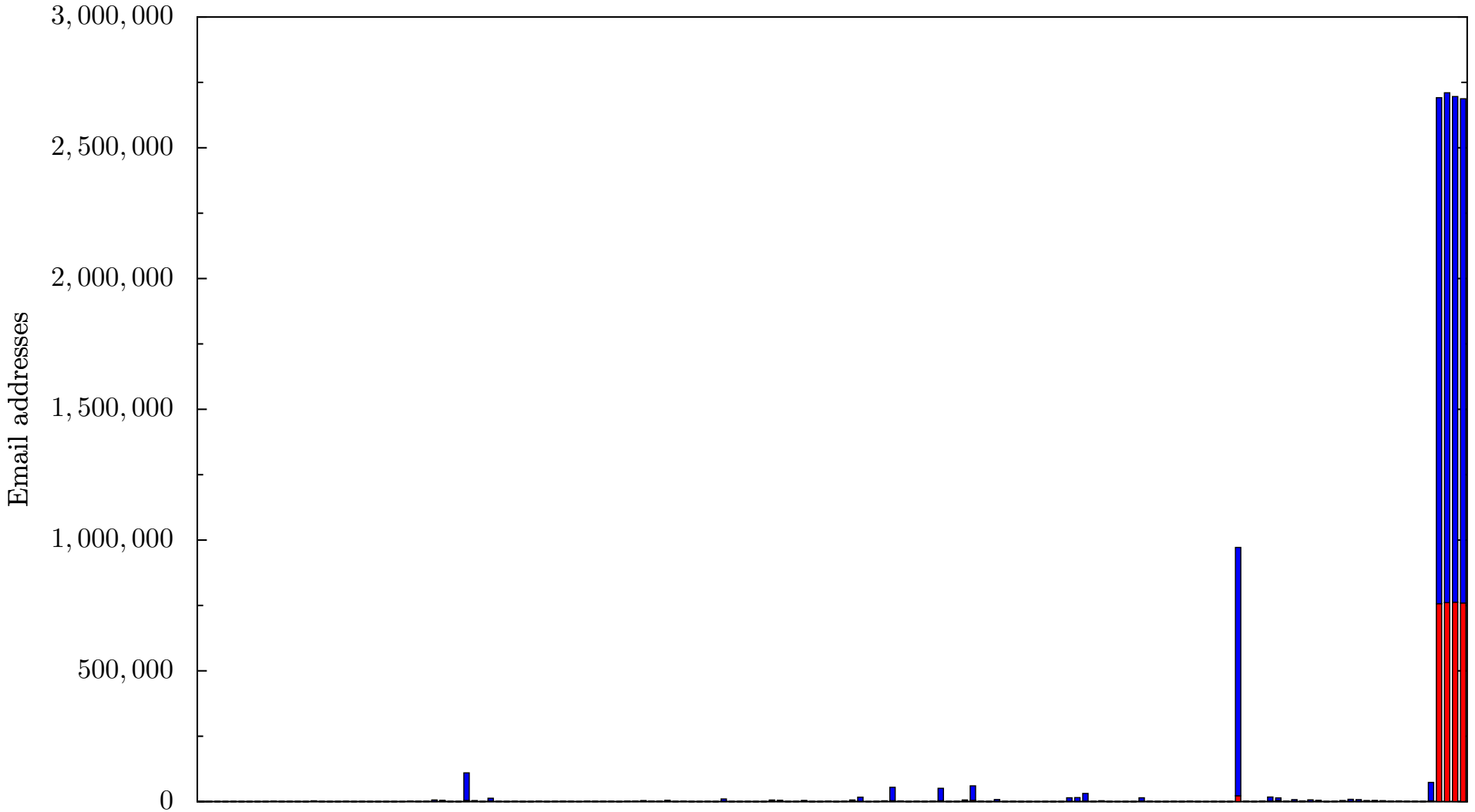
- Automatically building stop lists
- Hot drive identification

Automatic "stop lists:"
features on many drives are not pseudo-unique.

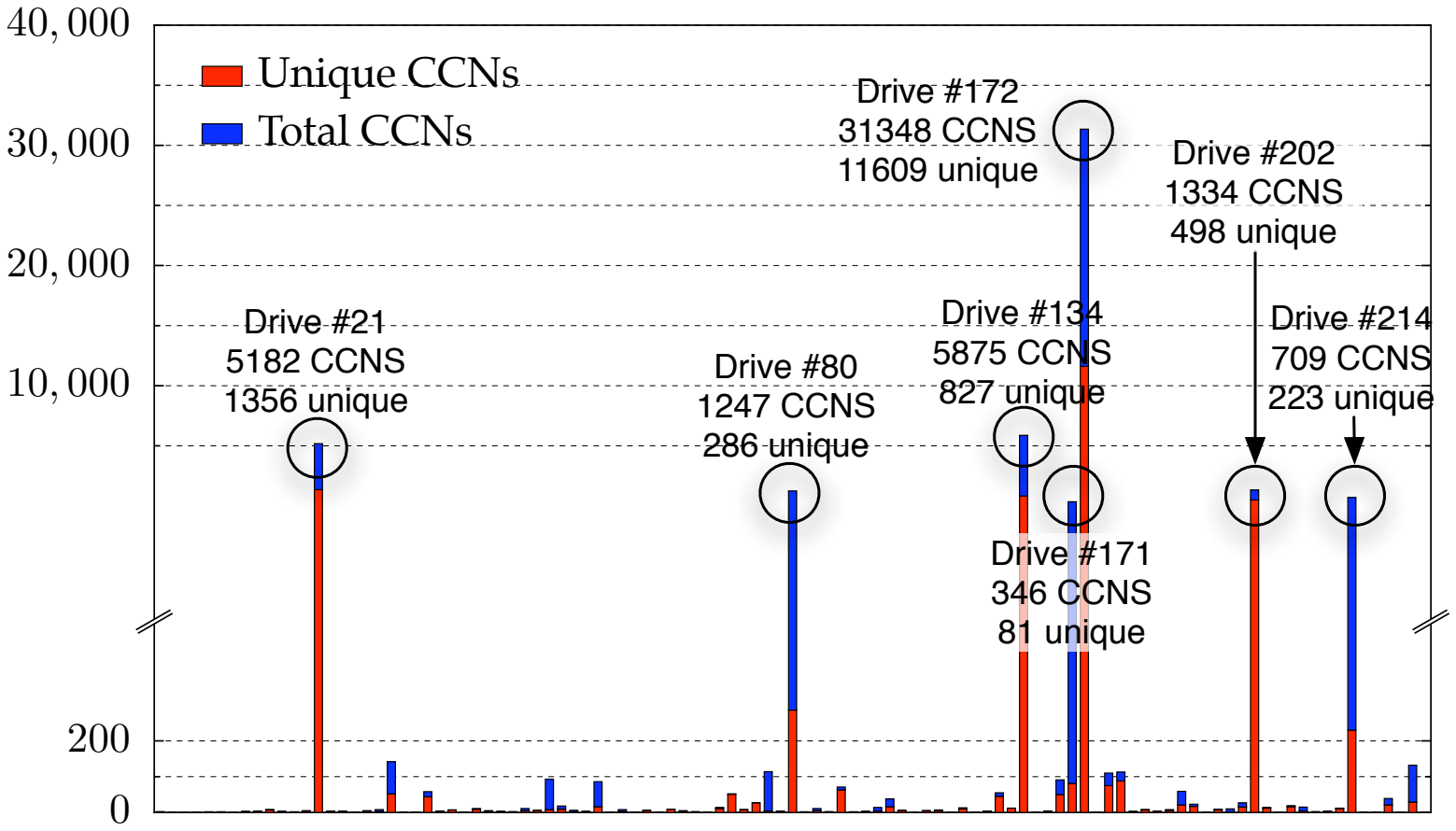
Extracted Email Address	Drives with address	Total count in corpus
CPS-requests@verisign.com	286	64424
server-certs@thawte.com	278	32873
premium-server@thawte.com	278	31141
Mouse.Exe@Mouse.Com	262	493
LMouse.Exe@LMouse.Com	262	493
personal-premium@thawte.com	253	14660
personal-freemail@thawte.com	250	14843
personal-basic@thawte.com	250	14290
inet@microsoft.com	244	31456
mazrob@panix.com(*)	221	3265
java-security@java.sun.com	200	1200
java-io@java.sun.com	198	413
someone@microsoft.com	195	6193
bugs@java.sun.com	192	351
ca@digsigtrust.com	173	36800
name@company.com	169	1763

*`mazrob@panix.com` **appears in** `clickerx.wav` **(Utopia Sound Scheme)**

A graph of # email addresses on each drive automatically identified drives used by bulk e-mailers.

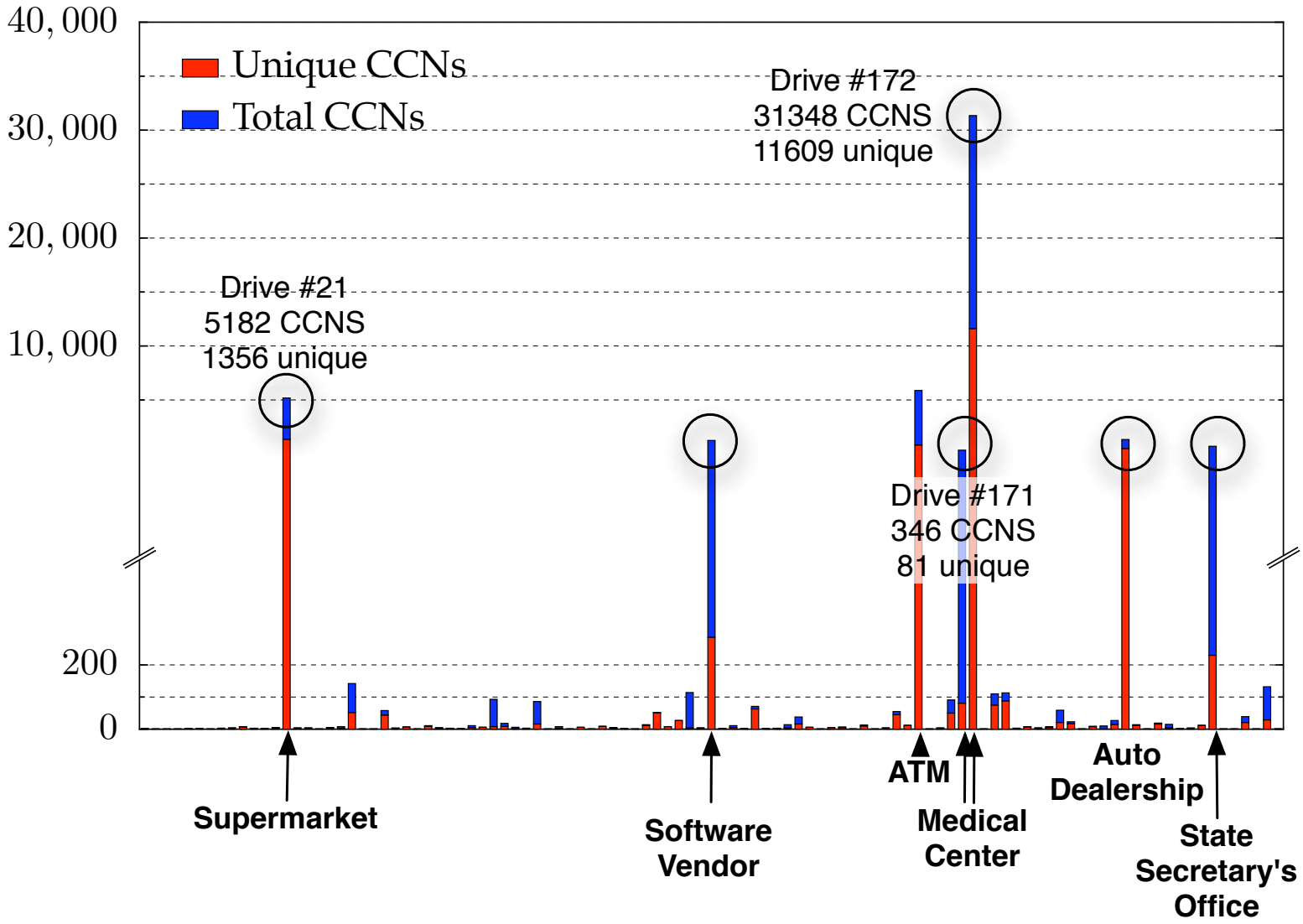


Hot drive identification:
Drives with high response warrant further attention.



Only 7 drives had more than 300 credit card numbers.

Hot drive identification:
Drives with high response warrant further attention.



These drives represent significant privacy violations.

First order analysis of # SSNs

Drive	Unique SSNs	Total SSNs
Drive #959	260	447
Drive #974	178	674
Drive #696	33	872
Drive #969	33	33
Drive #690	8	14
Drive #680	2	4

Drive #959 contained consumer credit applications.

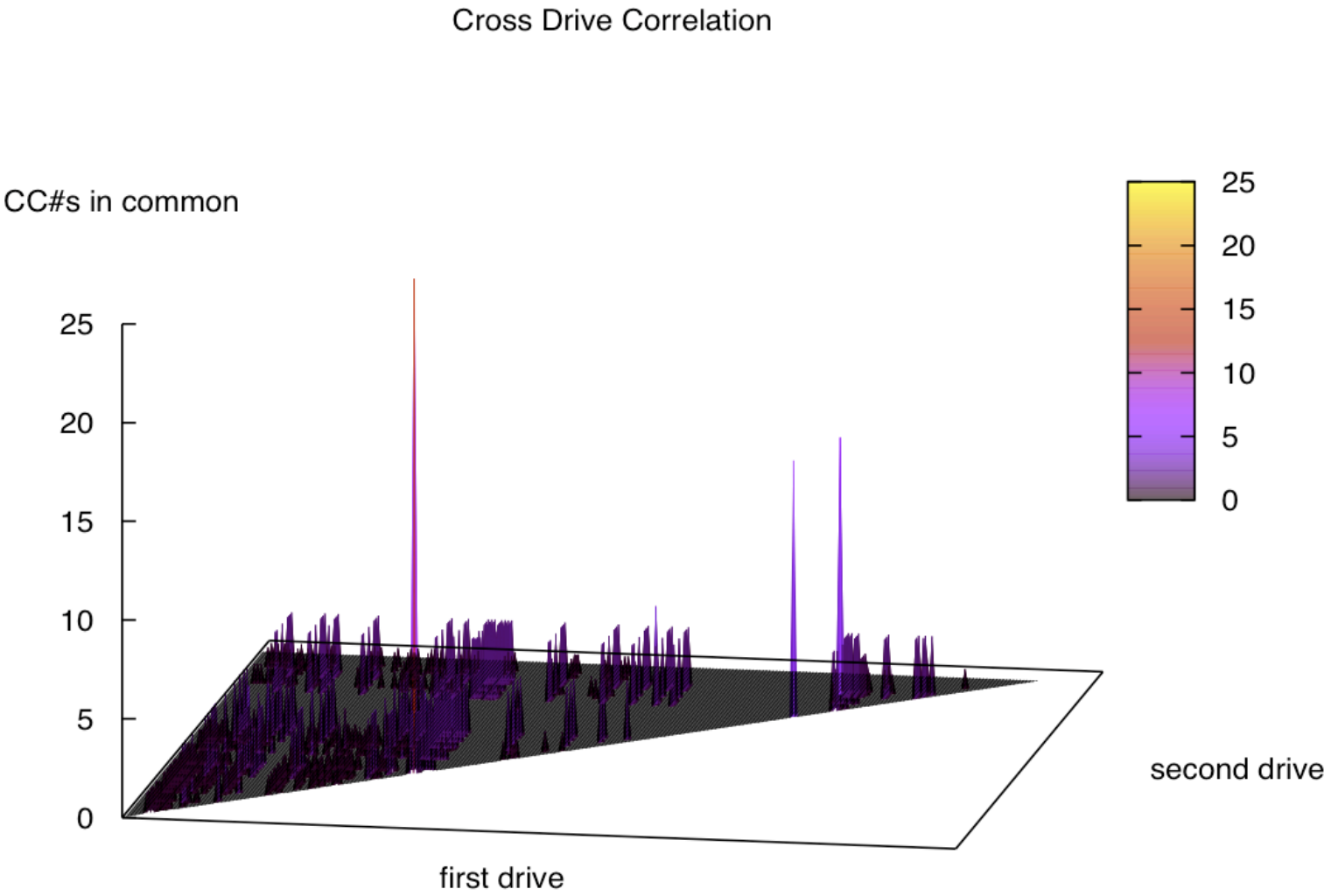
Second-order analysis uses the *multi-drive correlation*

$$
\begin{aligned}
D &= \#\text{ of drives} \\
F &= \#\text{ of extracted features} \\
d_0 \ldots d_D &= \text{Drives in corpus} \\
f_0 \ldots f_F &= \text{Extracted features} \\
FP(f_n, d_n) &= \begin{cases} 0 & f_n \text{ not present on } d_n \\ 1 & f_n \text{ present on } d_n \end{cases}
\end{aligned}
$$

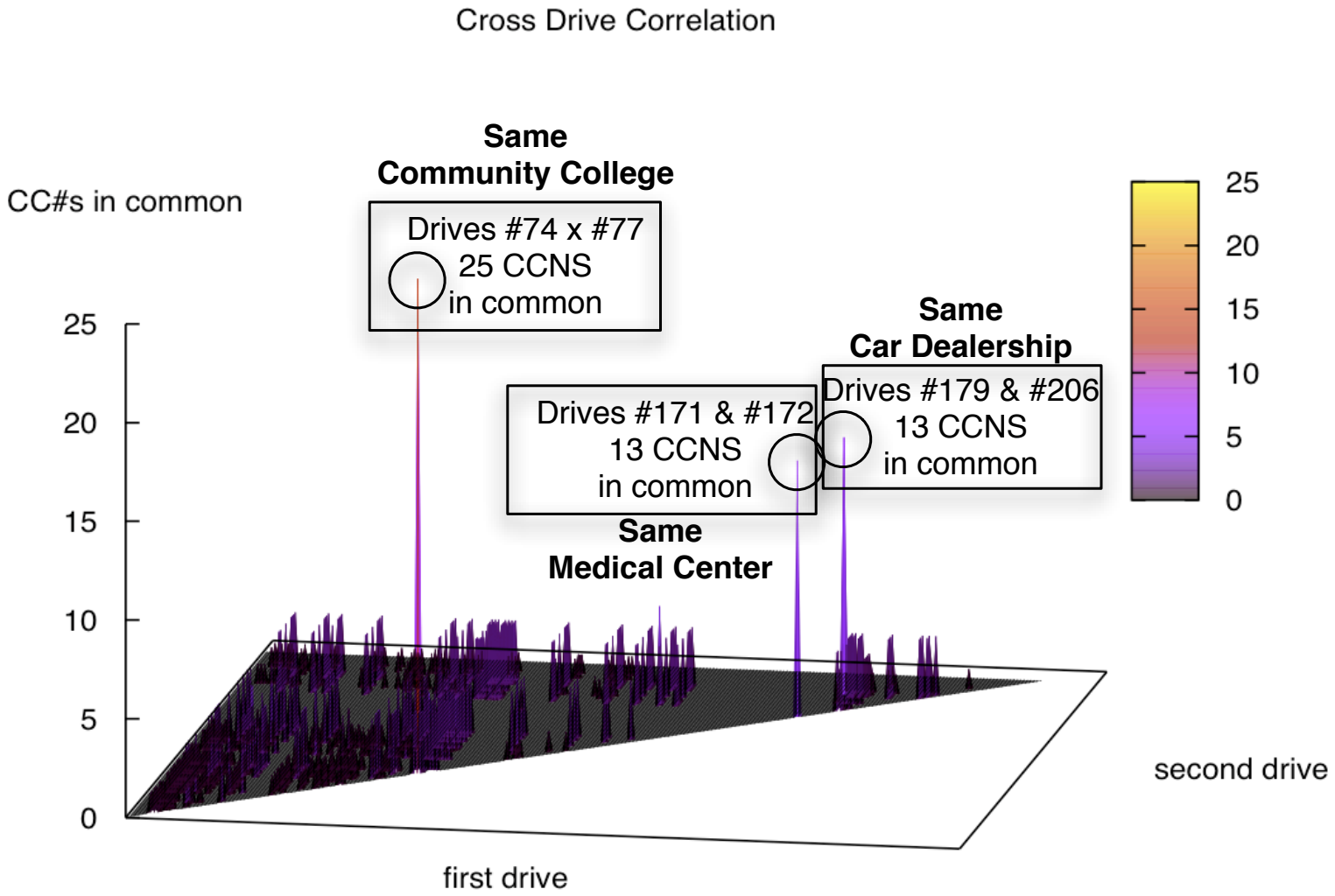
Scoring Function:

$$S_1(d_1, d_2) = \sum_{n=0}^{F} FP(f_n, d_1) \times FP(f_n, d_2)$$

Graph of scoring function:

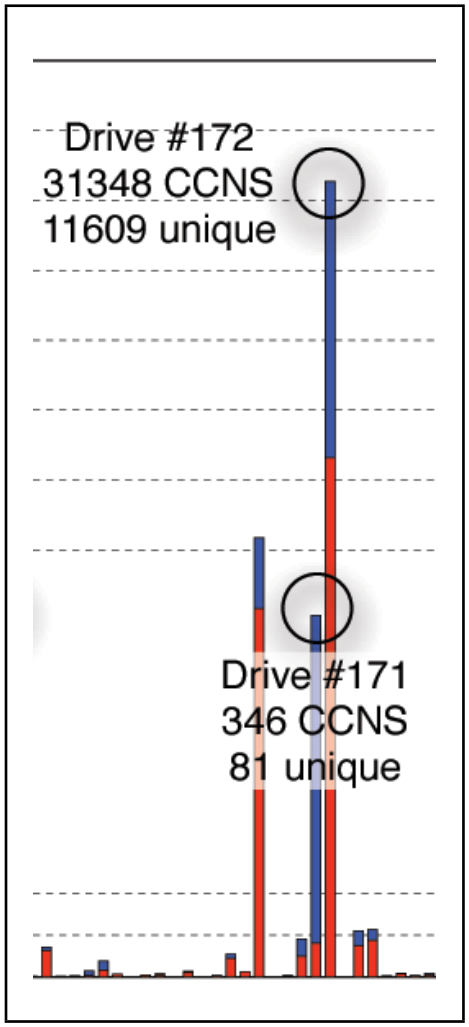


Graph of scoring function:



The three correlated drives have an extrinsic relationship. (180 drive corpus)

The correlation between Drives #171 and #172 tells a story...



Drive #171: Development drive

- Has source code.
- 346 CCNS; 81 unique.

Drive #172: Production system.

- 31,348 CCNS; 11,609 unique
- Oracle database (hard to reconstruct).

...The programmers used live data to test their system.

Other CCN correlations

#74, #77	Same college in Pacific Northwest. Correlated on CCN “false positive.”
#339 – #356	All used by same New York travel agency
#716, #718	Both from Union City, CA dealer
#814, #820	Both from same Stamford, CT dealer

In two cases, cross-drive correlation discovered drive cataloging errors!

SSN correlation: identical documents on different drives

SSN_1	#342, #343, #356	“Thanks, Laurie” memo
SSN_2	#350, #355	“great grandchildren” memo

But ignore these numbers:

666-66-6666	#313, #427, #429, #430, #612, #627, #744, #770, #808
123-45-6789	#328, #343, #345, #350, #351, #700
555-55-5555	#612, #690

Possible reasons for the same SSN found on two drives

- Two copies of the same document
- Two documents about the same person
- Accidental mismatch

Chance of a false match is 1 in 10^9.

Future Work 1: What is the best scoring function?

$$S_1(d_1, d_2) = \sum_{n=0}^{F} FP(f_n, d_1) \times FP(f_n, d_2)$$

Discount features that appear on many drives

$$
\begin{aligned}
DC(f) &= \sum_{n=0}^{D} FP(f, d_n) \\
&= \#\text{ of drives with feature } f
\end{aligned}
$$

$$
S_2(d_1, d_2) = \sum_{n=0}^{F} \frac{FP(f_n, d_1) \times FP(f_n, d_2)}{DC(f_n)}
$$

Weigh features that are rare on some drives, but high on others

$$
\begin{aligned}
DC(f) &= \#\text{ of drives with feature } f \\
FC(f,d) &= \text{count of feature } \mathit{f} \text{ on drive } \mathit{d}
\end{aligned}
$$

$$
S_3(d_1,d_2) = \sum_{n=0}^{F} \frac{FC(f_n,d_1) \times FC(f_n,d_2)}{DC(f_n)}
$$

More Future Work:

- Scaling cross-drive correlation to 10,000 drives.
- More sophisticated feature extraction based on Sleuth Kit.
- Use of sector hashes (MD5) to find fragments of documents on different drives.
- Combining CDA with carving and time line analysis.
- Automatically sanitize personal information for publication.

Acknowledgments

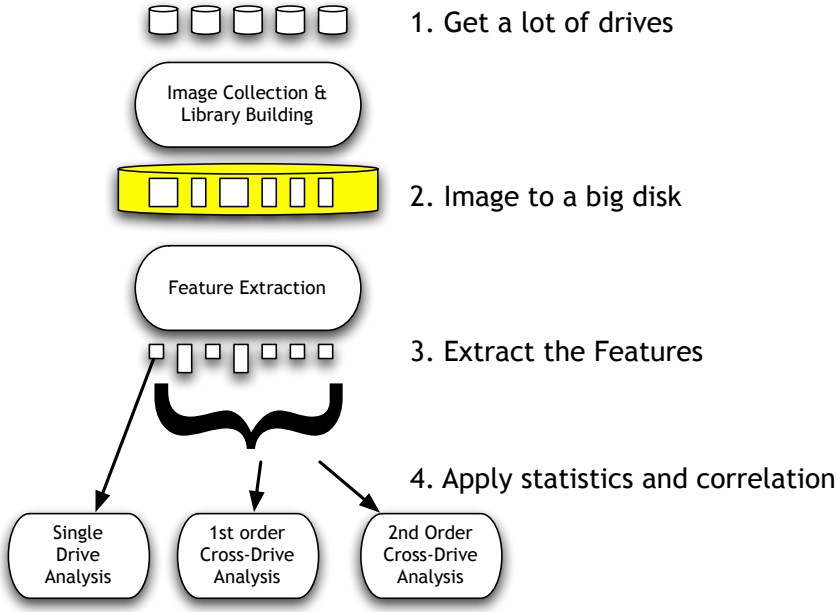
- Abhi Shelat (CCN) and Ben Gelb (email)
- Steve Bauer, Gene Spafford, Brian Carrier
- Basis Technology
- University of Auckland
- Harvard University CRCS

Summary

Large-scale forensics is an important problem

Feature Extraction and Cross-drive analysis allow:

- Better single-drive tools
- Intelligent stop-lists
- Identification of social networks



Questions?