Fast Contraband Detection In Large Capacity Disk Drives

By

Phil Penrose, William Buchanan and Richard Macfarlane

Presented At

The Digital Forensic Research Conference

DFRWS 2015 EU Dublin, Ireland (Mar 23rd- 26th)

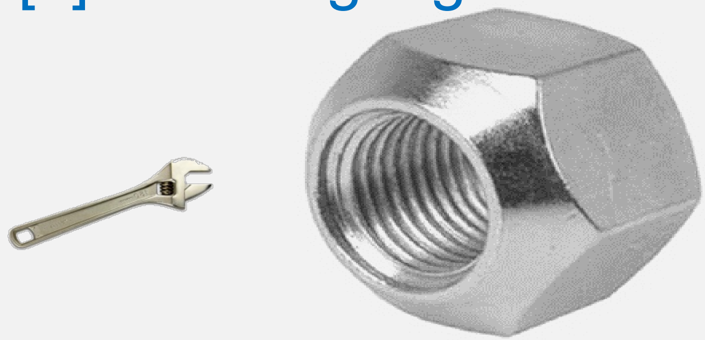
Edinburgh Napier UNIVERSITY

Presenter	Phil Penrose Research Student Edinburgh Napier University
Paper	Fast contraband detection in large capacity disk drives
Structure	Background to and motivation for this research Review of existing solutions The theory behind our design Experiments and results Questions

THE INDEPENDENT WEDNESDAY 04 FEBRUARY 2015
NEWS VIDEO PEOPLE VOICES SPORT TECH LIFE PROPERTY ARTS + ENTS TRAVEL
UK / World / Business / People / Science / Environment / Media / Technology / Education / Images /
News > UK > Crime
No evidence a … child porn inquiry who 'ki…
Napier UNIVERSITY
SEMPER VIGILO
POLIC…
child porn
mits suicide
arrested
THE SHIELDS Gazette
Local News Crime Politics Business Campaigns Education Health Regional News National News
Internet child-porn pics man considered suicide
DAILY NEWS
NEW YORK
SPORTS
ENTERTAINMENT
Sen. Lamar Alexa… commits suicide
Ryan Loskarn, former chief of staff to … himself in his parents' home in Marylan… possessing and distributing child pornog…
Follow
NEW YORK DAILY NE…
Mirror
CELEBRITY BIG BROTHER
All the latest news and gossip from the house
Most read
Live feeds
Top Videos
TRENDING
GENERAL ELECTION 2015 ANNE KIRKBRIDE FGM BLUE MONDAY EDUCATION ANONYMOUS
News Politics Football Celebs TV & Film Weird News
Sport Technology Money Travel Motoring
SUBSCRIPTIONS
Sign up now for 25% OFF
News · UK News · Gloucestershire Police
Policeman found dead in suspected suicide was child porn suspect
Los Angeles Times
LOCAL CALIFORNIA SPORTS ENTE…
YOU ARE HERE: LAT Home → Collections → Trends
Advertisement
COLUMN ONE
Child Porn Raids Lead to Suic…
Suspects may now end up killing themselves, expert…

Kryder’s Law - the areal density has been increasing at 40% per year and is nowhere near fundamental limits.

Garfinkel [1] - existing digital forensic tools do not scale



Roussev, Quates and Martell [2] - acquisition of a fast 3 TB hard disk - over 11 hours.

11 hours

Triage - a fast initial scan by sampling a digital device, conducted perhaps under severe time and resource constraints, to prioritise the device for possible further detailed investigation

- Be 99.9% accurate
- Give results in a reasonable time
- Execute on low specification legacy equipment.

Sampling – not scanning

Clusters not files

Files - Access slow
Relies on file system metadata so
no deleted files, partitions or unallocated space

Clusters – Sample clusters from the whole address
space of the disk. Sorted in order thus allowing a
sequential pass over the disk.
File system agnostic

RAM based reference data set

Existing ‘Triage’ Solutions

Kludge4

triage-ir

Internet history, the registry, file metadata, recently used files, image files……file hashing and lookup.
Disk imaging!

“freeing forensic analysts from the routine task of acquiring forensic evidence” Casey et al. [3]

Using sdhash similarity digests rather than simple hashing, (Roussev et al. [4])
“Similarity digests could only be used in the field in a selective manner, e.g., using a reference database of up to 1 GB.”

bulk extractor

Uses optimised database *hashdb* – specifically designed for the purpose.
Can do random sampling.

Reference Data Set (Contraband)

Police Scotland Child Pornography Image Database

- 5.1 million images Category 1 (the most serious)
- 1 million images Categories 2 to 5

Average image size 100 KiB =>25 × 4 KiB clusters

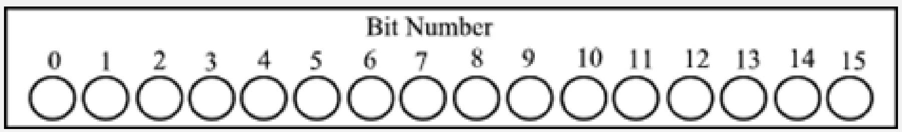
6 million × 25 cluster hashes = 150 million hashes

MD5 hash - 128 bits = 16 bytes

150 million × 16 bytes = 2.4 GB !!!!

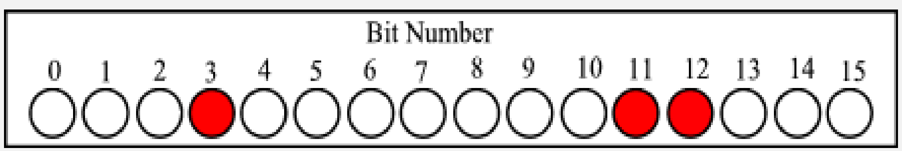
Bloom Filters

1. A bit array
2. A set of hash functions

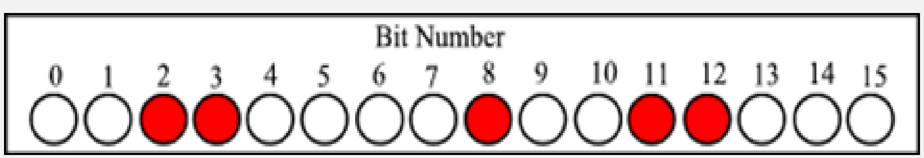


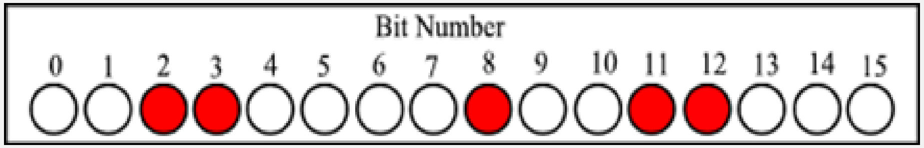
$h_1()$, $h_2()$, $h_3()$: <any input> → [0, 1, 2, ……….,15]

h_1(‘Jupiter’) = 3, h_2(‘Jupiter’) = 12, h_3(‘Jupiter’) = 11



h_1(‘Venus’) = 11, h_2(‘Venus’) = 2, h_3(‘Venus’) = 8





Lookup -
h_1('Saturn') = 1, h_2('Saturn) = 3, h_3('Saturn') = 8

h_1('Mars') = 11, h_2('Mars') = 3, h_3('Mars') = 8 !!

Values important to the design of our Bloom filter:

m = the number of bits in the array which represents the Bloom filter.
Initially all bits are set to 0
n = the number of elements added to the filter
k = the number of independent hash functions h_1, h_2, ….. , h_k used
p = the probability of a false positive

Since $m = 2^4$ note that 4 bits are required for each hash

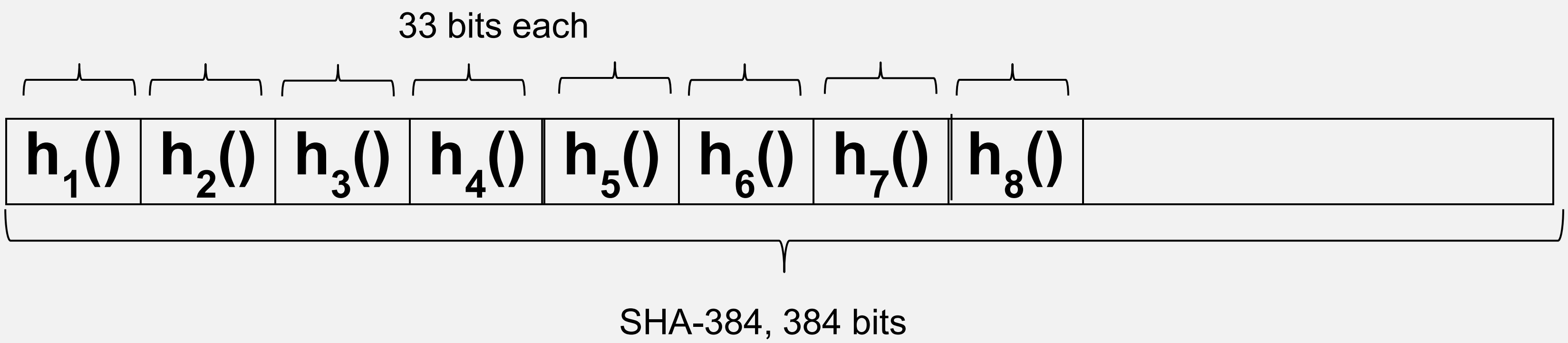
$$p = P(false\ positive) \approx \left(1 - e^{-\frac{kn}{m}}\right)^k$$ (Mitzenmacher and Vadhan [5])

False positive probabilities for varying values of m and k (number entries in filter = 200 million)

	Filter size (MiB)					
k	512	600	700	800	900	1024
4	0.000834149	0.000466407	0.000263157	0.000159487	0.000102189	0.000062537
6	0.000209786	0.000091112	0.000039869	0.000019271	0.000010073	0.000004912
8	0.000087538	0.000030242	0.000010473	0.000004100	0.000001769	0.000000696
10	0.000051133	0.000014373	0.000004016	0.000001292	0.000000466	0.000000149
12	0.000037893	0.000008855	0.000002034	0.000000546	0.000000166	0.000000044
14	0.000033433	0.000006629	0.000001274	0.000000289	0.000000075	0.000000017
16	0.000033607	0.000005763	0.000000942	0.000000183	0.000000041	0.000000008

1 GiB array = 2^{30} bytes = 2^{33} bits
So 33 bit hashes are required

SHA-384 algorithm generates a 384 bit hash
We ‘slice off’ 8 × 33 bit independent hashes

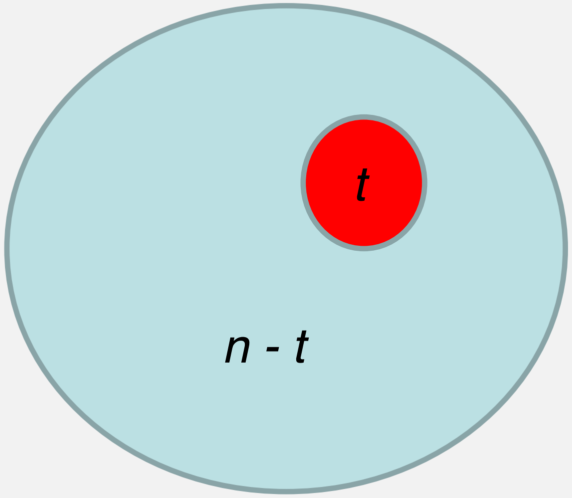


Calculating sample size:

Disk size	n clusters
Target size	t clusters
Sample size	k clusters

Probability of a 'hit' = p = 99.9%

Sampling k items from n with no replacement – The Urn Problem - Hypergeometric Distribution



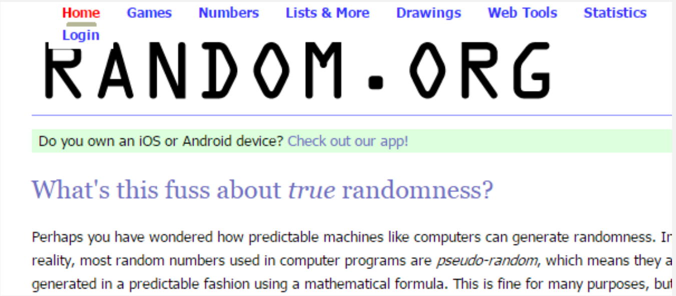
Target size 4 MiB – Probability of at least one target cluster in sample

Disk Size (GB)	Sample Size: 100000	200000	300000	400000	500000	600000	700000	800000	900000	1000000
120	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
250	0.79	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
320	0.71	0.91	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00
500	0.54	0.79	0.90	0.96	0.98	0.99	1.00	1.00	1.00	1.00
1000	0.32	0.54	0.69	0.79	0.86	0.90	0.94	0.96	0.97	0.98

Target size 20 MiB – Probability of at least one target cluster in sample

Disk Size (GB)	Sample Size: 100000	200000	300000	400000	500000	600000	700000	800000	900000	1000000
120	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
320	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
500	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1000	0.86	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

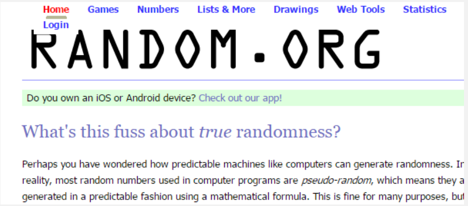
Edinburgh Napier
UNIVERSITY



Add 200 million random clusters to filter

1 million original random clusters – all reported present

Test

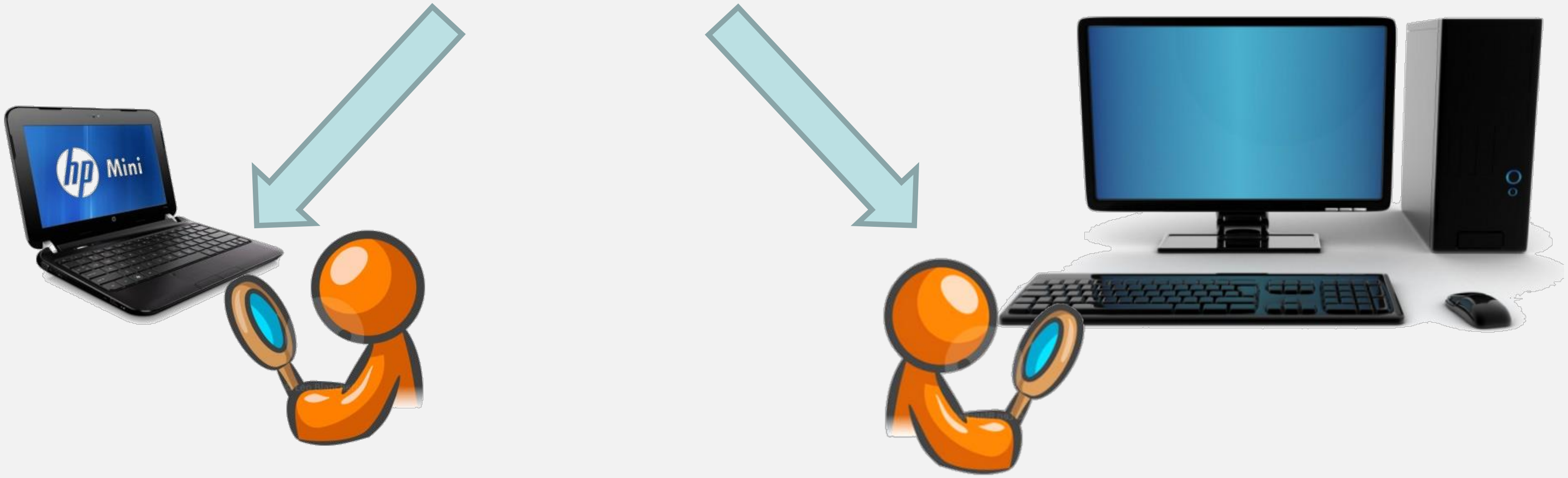


100 MB real images added to filter

1 million fresh random clusters – 1 false positive

Testing

4 MB of random images
from real image set
added to each test drive



Sampled for contraband

Edinburgh Napier UNIVERSITY

Core i3 Desktop PC Sampling Accuracy and Speed

Disk and Size	Target Size	Samples	Hits	False Positives	Time min:sec
250 GB SSD	4 MB	416700	31	0	00:42
250 GB SSD	20 MB	79700	7	0	00:07
250 GB USB HDD	4 MB	416700	7	0	35:02
250 GB USB HDD	20 MB	79700	4	0	08:31
1 TB HDD	4 MB	1666600	6	1	108:54
1 TB HDD	20 MB	318900	8	2	27:42

Intel Atom Netbook Sampling Accuracy and Speed

Disk and Size	Target Size	Samples	Hits	False Positives	Time min:sec
250 GB SSD	4 MB	416700	8	0	09:33
250 GB SSD	20 MB	79700	12	0	01:54
250 GB USB HDD	4 MB	416700	5	0	44:34
250 GB USB HDD	20 MB	79700	7	0	11:04

Edinburgh Napier UNIVERSITY

Triage –
a fast initial scan by sampling a digital device✓

- Be 99.9% accurate ✓
- Give results in a reasonable time ✓
- Execute on low specification legacy equipment ✓

However -

Edinburgh Napier UNIVERSITY

We have used random data for initial testing which eliminated false positives due to <u>non-distinct blocks</u>.

Real data will have non-distinct blocks (Young et al. [6]).
Initial testing shows possibly 1% of blocks could be common.
This problem is currently being addressed by Garfinkel et al..

In the mean time we read two blocks for every sample.
If a hit – process block 2.
If both hits – it's a hit, otherwise false positive.

Sector alignment had not proved a problem. Some researchers have read two disk clusters at a time then do 8 filter lookups instead of one – one for each sector offset.
We have successfully used wmic.exe

Thankyou.

References

[1] S. Garfinkel, “Digital forensics research: The next 10 years,” Digital Investigation, vol. 7, pp. S64–S73, Aug. 2010.

[2] V. Roussev, C. Quates, and R. Martell, “Real-time digital forensics and triage,” *Digital Investigation*, vol. 10, no. 2, pp. 158 – 167, 2013.

[3] E. Casey, G. Katz, and J. Lewthwaite, “Honing digital forensic processes,” *Digital Investigation*, vol. 10, no. 2, pp. 138–147, Sep. 2013.

[4] V. Roussev, C. Quates, and R. Martell, “Real-time digital forensics and triage,” *Digital Investigation*, vol. 10, no. 2, pp. 158 – 167, 2013.

[5] M. Mitzenmacher and S. Vadhan, “Why simple hash functions work: exploiting the entropy in a data stream,” *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 746–755, 2008.