Building and decaying a file corpus for sub-sector analysis

By:
Dominique Calder (George Mason University)

From the proceedings of
The Digital Forensic Research Conference
DFRWS USA 2022
July 11-14, 2022

DFRWS 2022 VIRTUAL USA

Building And Decaying A File Corpus For Sub-sector Analysis

Dominique Calder, PhD Student
Dr. James Jones
George Mason University

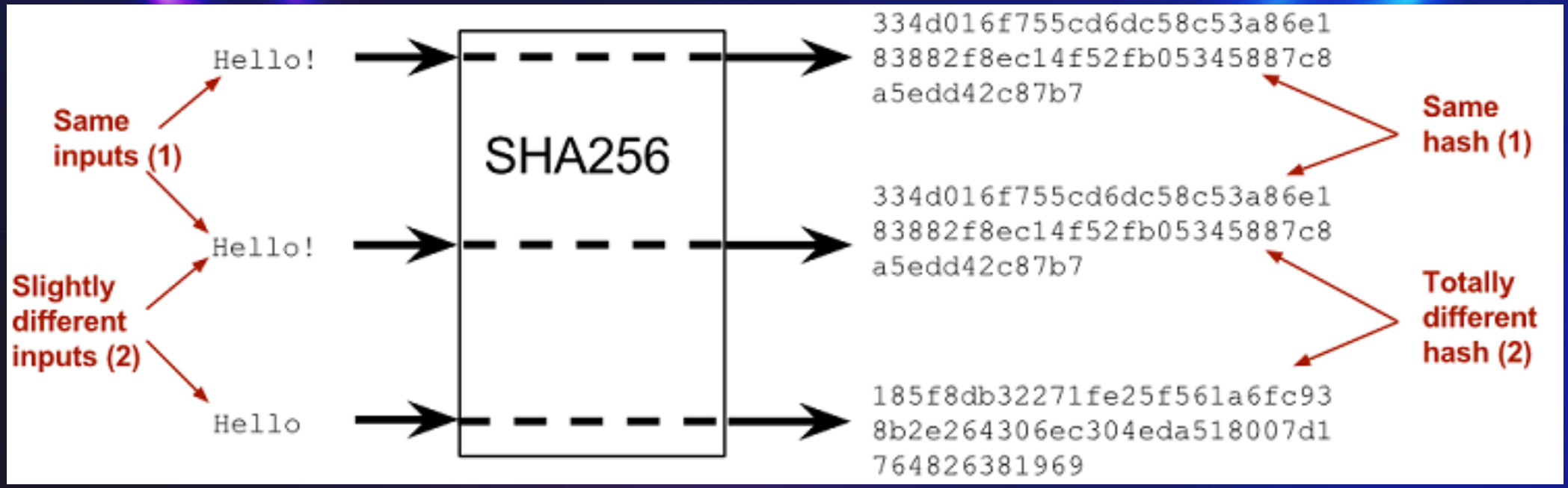
- Law Enforcement and private organizations process massive amounts of digital evidence (up to petabytes) and need to quickly determine if data of interest is present on a device.
- Traditional file search methods are time intensive and yield the best results when an exact match is identified.
- Files that have been partially overwritten go undetected when searching via file hashes.

BACKGROUND

DFRWS 2022 VIRTUAL USA

- File hashes verify the integrity of a file. The slightest change to the file changes the file hash completely.
- This integrity tracking mechanism (file hashing) makes it difficult, nearly impossible to track files that are largely similar, but share some differences.
- Files and artifacts of interest often go unfound during investigations when using file hashes as the sole source of integrity matches, leaving the narrative only partially understood.

PROBLEM


Hello!
Same inputs (1)
Hello!
Slightly different inputs (2)
Hello
SHA256
334d016f755cd6dc58c53a86e1
83882f8ec14f52fb05345887c8
a5edd42c87b7
Same hash (1)
334d016f755cd6dc58c53a86e1
83882f8ec14f52fb05345887c8
a5edd42c87b7
Totally different hash (2)
185f8db32271fe25f561a6fc93
8b2e264306ec304eda518007d1
764826381969

PROBLEM

- Sector hashes are unique digests of data that are more granular than individual file hashing.
- Using a public document corpus [govdocs1 = 1 million files], we propose to generate sector hashes for each file.
- This microscopic view creates a many-to-one correlation for each file with the associated sectors and enables forensic examiners to understand the data more granularly to make more accurate determinations in file activity.
- These research efforts aid in potentially identifying criminal networks along with detecting anti-forensic and data tampering tactics.

PROPOSED RESEARCH

DFRWS 2022 VIRTUAL USA

Goal: Develop a scalable algorithm that can accurately infer the past presence of a file given arbitrary sectors

1. Sector hash ingested data blocks (with sliding window)
2. Select statistically improbable features/feature combinations to search for in sector hash blocks.
3. Analyze sector hashes in small chunks via random sector sampling for matches
4. Determine threshold criteria to determinate evaluation metrics for true/false positive and non-classification detection (indeterminate)
5. Determine matches based on entropy score and distribution, relative order, and similarity
6. Iterate algorithm with artificial file decay increased by 10%

METHODOLOGY

DFRWS 2022 VIRTUAL USA

10 sector hashes generated for a 40KB file
3460 hashes generated for a 18MB file

Sector Hash (MD5)	File	Offset
35f3d37ea3799e164f7da7186d9d7903	D:\doc\000\000001.doc	0-4095
f5cf541c9f22a2cc98d2db37313d2616	D:\doc\000\000001.doc	4096-8191
cfe5fd285485d4b76190ad1c37993f08	D:\doc\000\000001.doc	8192-12287
71a676aa4fab086d2396b8a30c52f775	D:\doc\000\000001.doc	12288-16383
cd650a8ce27f32096df6e01e055f2d00	D:\doc\000\000001.doc	16384-20479
ffc62f2fce9a315389d497cf4d64f2f5	D:\doc\000\000001.doc	20480-24575
7f2e2e51253dd6e67ca170ea3daae350	D:\doc\000\000001.doc	24576-28671
54c03f91551ae88aef04f7bdd704f03e	D:\doc\000\000001.doc	28672-32767
a8d1cb7e6b6c36b80e8a945413a987cd	D:\doc\000\000001.doc	32768-36863
f900411b0c88c739c4bc3d794a7064d8	D:\doc\000\000001.doc	36864-40959

METHODOLOGY

DFRWS 2022 VIRTUAL USA

Sector Hash (MD5)	File	Offset	Decay0	Decay1	Decay2	Decay3	Decay4	Decay5	Decay6	Decay7
35f3d37ea3799e164f7da7186d9d7903	D:\doc\000\000001.doc	0-4095	0	0	0	0	0	0	0	0
f5cf541c9f22a2cc98d2db37313d2616	D:\doc\000\000001.doc	4096-8191	0	0	0	0	0	0	0	0
cfe5fd285485d4b76190ad1c37993f08	D:\doc\000\000001.doc	8192-12287	0	0	0	0	0	0	0	0
71a676aa4fab086d2396b8a30c52f775	D:\doc\000\000001.doc	12288-16383	0	0	0	0	0	0	0	0
cd650a8ce27f32096df6e01e055f2d00	D:\doc\000\000001.doc	16384-20479	0	0	0	0	0	0	0	0
ffc62f2fce9a315389d497cf4d64f2f5	D:\doc\000\000001.doc	20480-24575	0	0	0	0	0	0	0	0
7f2e2e51253dd6e67ca170ea3daae350	D:\doc\000\000001.doc	24576-28671	0	0	0	0	1	1	1	1
54c03f91551ae88aef04f7bdd704f03e	D:\doc\000\000001.doc	28672-32767	0	0	0	0	1	1	1	1
a8d1cb7e6b6c36b80e8a945413a987cd	D:\doc\000\000001.doc	32768-36863	0	0	1	1	1	1	1	1
f900411b0c88c739c4bc3d794a7064d8	D:\doc\000\000001.doc	36864-40959	0	1	1	1	1	1	1	1

Example of sector hashes after decay algorithm implementations

METHODOLOGY

- Files from govdocs1 have been hashed at the sector level and combined into a database for decay analysis.
- Decay iterations have started to determine how files persist after operating system events such as deletions and file overwrites.
- Analysis is still on-going to include how entropy of the files change during decay iterations.

RESEARCH: PRESENT STATE

DFRWS 2022 VIRTUAL USA

- We plan to use this research as grounds to determine how data persists on traditional hard drive (HDD) and solid-state drives (SSD).
- Ultimately, this research will help find matches between diverse and dispersed data sets using sector hashing.

RESEARCH: FUTURE STATE

DFRWS 2022 VIRTUAL USA

REFERENCES

1. Roussev, V. Scalable data correlation. International Conference on Digital Forensics (IFIP WG 11.9) (January 2012)
2. Zanganeh, Omid, et al. “Partial Fingerprint Alignment and Matching Through Region-Based Approach.” Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia, 2015, pp. 1–10., doi:10.1145/2837126.2837132.
3. Simson Garfinkel, Alex Nelson, Douglas White, Vassil Roussev, “Using purpose-built functions and block hashes to enable small block and sub-file forensics,” in Digital Investigation, Volume 7, Supplement, 2010, Pages S13-S23.
4. Roussev V. (2010) Data Fingerprinting with Similarity Digests. In: Chow KP., Shenoi S. (eds) Advances in Digital Forensics VI. (2010)