Known Hash Filtering: An Efficient Way to Exclude Irrelevant Files and Display Files of Interest in Digital Examinations

Jessica A. Smith*, B.S.[1]; Lyndsay Haak, CASA, CCE, ACT, MCFE, MCGE, MCVE[2]; Timothy Suggs[2], James Trevillian (Ret.)[2], Josh Brunty, Sc.D[1]

[1]Marshall University Forensic Science Center, 1401 Forensic Science Drive, Huntington, WV 25701

[2]North Carolina State Crime Laboratory, 121 E Tryon Road, Raleigh, NC 27603

Abstract

Hash sets are collections of data that are compiled of hash values, or unique digital fingerprints, that match known files. Such hash sets include the Reference Data Sets (RDS) from the National Software Reference Library (NSRL) and Project VIC. The goals of this project are to use the NSRL sets to hide known irrelevant files, import Project VIC hash sets, create custom hash sets to simulate locating known files of interest, and determine if mobile devices change the hash value of known images when saved to a device. Laboratory procedures were created outlining the steps to obtain these goals. Part One resulted in five of the six programs successfully filtering files, decreasing the number of files. Part Two resulted in three of the six programs successfully importing Project VIC hash sets, and five of the six programs successfully located known files of interest. In Part Three, hash values of known files were subject to change when saved on iOS devices while Android devices did not change hash values when saved to the device.

Introduction

- Validated processes, such as known hash set filtering, have been developed to reduce a portion of the files examined.
- The purpose of known hash set filtering is to hide known irrelevant files and locate known files of interest.
- The National Software Reference Library (NSRL) is the largest database of known hash values, and includes Reference Data Sets (RDS) that consist of the following file information:
  - *Name*
  - *Size*
  - *Operating System*
  - *Type of software associated with the OS*
- The NSRL is only representative of known "good files" which includes a device's default files.
- Opposite to the NSRL, Project VIC is a collection of hash values of known "bad files" from known contraband.
- Project VIC utilizes PhotoDNA which creates a digital fingerprint, hash value, to determine image matches even when the original version is modified.
- This project focuses on the creation of procedures utilizing validated forensic tools and known hash set filtering to determine if the process efficiently aids in digital forensic examinations for state crime laboratories.
- The parts of this project include:
  - *Part One: Filtering Irrelevant Files using NSRL*
  - *Part Two: Project VIC & Custom Hash Sets*
  - *Part Three: Hashes & Mobile Devices*

Materials

Table 1 Digital Forensic Software Programs

Autopsy®	Forensic ToolKit (FTK®)	Open Text™ EnCase™ Forensic
Magnet Axiom™ Process and Examine	Cellebrite Inspector	Griffeye Analyze DI

Table 2 Additional Project Materials

Forensic Computers, Inc. forensic computer tower	Test Images (Table 3)	Project VIC hash sets (Table 4)	NSRL hash sets (Table 5)
Apple iPhone 4s	Samsung Galaxy	iFit tablet	HashCalc

Table 3 Test Image Operating Systems

Test Image	Test Case 1	Test Case 2	Test Case 3	Test Case 4	Image XP
Operating System	Windows 7	Windows XP	Windows 10	Windows 10	Windows XP

Table 4 Project VIC hash sets with corresponding program and imported hashes

Software Program	Project VIC Hash Set Version	Number of Hashes
Forensic ToolKit (FTK®)	06.2023	7,230,200
Magnet Axiom™	06.2023	15,280,200
Cellebrite Inspector	2020.02.08 – 2020.05.06	12,579,316
Griffeye Analyze DI	06.2023	15,463,589

Table 5 NSRL hash sets with corresponding program, import time, and imported hashes

Software	FTK®	EnCase™ Forensic	Magnet Axiom™	Cellebrite Inspector	Griffeye Analyze DI
Imported Hash Set	Modern 2023.03.1	Modern Minimal 2023.06.1	Modern Minimal 2023.06.1	NSRLFile.txt 11.2022	Modern 2023.03.1
Hash Set Import Time	~ 20 hrs	~ 1 hrs 15 min	~ 2 hrs	8 – 10 min	> 24 hrs
Number of Hashes	752,729,249	62,512,020	62,512,061	54,438,893	753,729,249

Methodology

Part 1: Filtering Irrelevant Files using NSRL

- Determine if the NSRL hash sets can be ingested in the six digital forensic tools
- Test the ability to filter these known hash sets and hide irrelevant files
- Utilize five test cases to evaluate how the NSRL will interact with different systems
- Create laboratory procedures outlining the specific steps for each of the six digital forensic tools

Part 2: Project VIC & Custom Hash Sets

- Determine whether all six programs could import a Project VIC hash set version
- Create two custom hash sets to simulate how each program could alert or display files of interest
- Import and apply the custom hash sets to two test cases to test each programs' ability to locate files of interest instead of hiding them

Part 3: Hashes & Mobile Devices

- Email images from the custom hash set to two mobile devices: Apple and Android
- Apple device: (1) save images directly to the Photos Library, (2) download to iCloud, then sync
- Android device: (1) save images directly to Gallery on the device, (2) save directly to Google Photos
- Using HashCalc, determine discrepancies with hash values of the devices' images to the known hashes

Results & Discussion

Table 6 FTK® file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	244,946	72,700	172,246	**29.68%**
Test Case 2	191,347	17,965	173,382	**9.39%**
Test Case 3	320,873	229,609	91,264	**71.56%**
Test Case 4	584,243	352,906	231,337	**60.40%**
Image XP	451,531	265,259	186,272	**58.75%**

Table 7 EnCase™ Forensic file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	140,560	65,487	75,073	**46.59%**
Test Case 2	85,325	16,822	68,503	**19.72%**
Test Case 3	248,872	149,684	99,188	**60.14%**
Test Case 4	491,804	275,663	216,141	**56.05%**
Image XP	350,065	184,742	165,323	**52.77%**

Table 8 Magnet Axiom™ file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	158,017	20,556	137,461	**13.01%**
Test Case 2	206,959	7,756	199,203	**3.75%**
Test Case 3	137,322	26,590	110,732	**19.36%**
Test Case 4	188,372	30,571	157,801	**16.23%**
Image XP	202,470	31,177	171,293	**15.40%**

Table 9 Cellebrite Inspector file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	143,389	63,129	80,260	**44.03%**
Test Case 2	100,577	14,018	86,559	**13.94%**
Test Case 3	143,534	84,893	58,641	**59.14%**
Test Case 4	344,836	181,318	163,518	**52.58%**
Image XP	224,733	107,858	116,875	**47.99%**

Table 10 Griffeye Analyze DI file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	102,338	64,776	37,562	**63.30%**
Test Case 2	82,361	18,093	64,268	**21.97%**
Test Case 3	167,202	135,856	31,346	**81.25%**
Test Case 4	307,990	232,076	75,914	**75.35%**
Image XP	234,058	157,620	76,438	**67.34%**

Table 17 Statistical analysis of program consistency with multimedia files filtered

Test Case	1	2	3	4	XP
FTK® Remaining Files	172,246	173,382	91,2641	231,337	186,272
EnCase™ Remaining Files	75,073	68,503	99,188	216,141	165,323
Magnet Axiom™ Remaining Files	137,461	199,203	110,732	157,801	171,293
Cellebrite Inspector Remaining Files	80,260	86,559	58,641	163,518	143,240
Griffeye DI Remaining Files	37,562	64,268	31,346	75,914	76,438
Average Remaining Files	100,520	118,383	78,234	168,942	143,249
Standard Deviation	48,023	56,544	29,146	54,642	40,704
Average Remaining Files (Axiom omit)	91,285	98,178	70,110	171,728	136,227
Standard Deviation (Axiom omit)	49,561	44,219	27,052	60,773	42,721
Low 1 SD (Axiom omit)	41,724	53,959	43,058	110,954	93,506
High 1 SD (Axiom omit)	140,846	142,397	97,162	232,501	178,948
Number of Programs Outside 1 SD (Axiom omit)	2	1	2	1	1
Low 2 SD (Axiom omit)	-7,837	9,741	16,606	50,181	50,785
High 2 SD (Axiom omit)	190,408	186,615	124,213	293,274	221,669
Number of Programs Outside 2 SD (Axiom omit)	0	0	0	0	0

Table 12 FTK® multimedia file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	7,087	4,064	3,023	**57.34%**
Test Case 2	37,828	3,610	34,218	**9.54%**
Test Case 3	52,300	48,707	3,593	**93.13%**
Test Case 4	56,223	51,402	4,821	**91.43%**
Image XP	53,365	50,533	2,832	**94.69%**

Table 13 EnCase™ Forensic multimedia file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	47,648	15,590	32,058	**32.72%**
Test Case 2	59,077	6,788	52,289	**11.49%**
Test Case 3	36,695	18,379	18,316	**50.09%**
Test Case 4	39,909	21,404	18,505	**53.63%**
Image XP	37,946	21,291	16,655	**56.11%**

Table 14 Magnet Axiom™ multimedia file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	6,013	3,284	2,729	**54.62%**
Test Case 2	35,772	2,364	33,408	**6.61%**
Test Case 3	17,692	14,279	3,413	**80.71%**
Test Case 4	24,331	20,474	3,857	**84.15%**
Image XP	22,271	17,453	4,818	**78.37%**

Table 15 Cellebrite Inspector multimedia file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	6,535	3,377	3,158	**51.68%**
Test Case 2	42,110	1,454	40,656	**3.45%**
Test Case 3	16,795	13,389	3,406	**79.72%**
Test Case 4	22,877	19,291	3,586	**84.32%**
Image XP	21,157	16,266	4,891	**76.88%**

Table 16 Griffeye Analyze DI multimedia file filter comparison

Test Case	Total Files	Filtered Files	Remaining Files	Percent Decrease
Test Case 1	6,614	3,436	3,178	**51.95%**
Test Case 2	37,495	3,367	34,128	**8.98%**
Test Case 3	51,560	47,926	3,634	**92.95%**
Test Case 4	56,771	50,792	5,797	**89.47%**
Image XP	54,221	49,859	4,362	**91.96%**

Table 17 Statistical analysis of program consistency with multimedia files filtered

Test Case	1	2	3	4	XP
FTK® Remaining Files	3,023	34,218	3,593	4,821	2,832
EnCase™ Remaining Files	2,729	33,408	3,413	3,857	4,818
Magnet Axiom™ Remaining Files	32,058	52,289	18,316	18,505	16,655
Cellebrite Inspector Remaining Files	3,158	40,656	3,406	3,586	4,891
Griffeye DI Remaining Files	3,178	34,128	3,634	5,979	4,362
Average Remaining Files	8,829	38,940	6,472	7,350	6,712
Standard Deviation	11,616	7,172	5,923	5,641	5,027
Average Remaining Files (Axiom omit)	3,022	35,603	3,512	4,561	4,226
Standard Deviation (Axiom omit)	179	2,934	103	939	830
Low 1 SD (Axiom omit)	2,843	32,668	3,408	3,622	3,396
High 1 SD (Axiom omit)	3,201	38,537	3,615	5,499	5,056
Number of Programs Outside 1 SD (Axiom omit)	1	1	2	2	1
Low 2 SD (Axiom omit)	2,663	29,734	3,305	2,683	2,566
High 2 SD (Axiom omit)	3,381	41,471	3,718	6,438	5,885
Number of Programs Outside 2 SD (Axiom omit)	0	0	0	0	0

Conclusions

Part One: Filtering Irrelevant Files using NSRL

- All programs, except Autopsy®, successfully performed the known hash set filtering process.
- There is no standard relating to the threshold needed to designate a percent decrease as "efficient"; therefore, 50% was determined as the threshold for this project.
  - Values less than 50%, in red text, were deemed as inefficiently decreased.
  - Values greater than or equal to 50%, in black text, were deemed as efficiently decreased.
- Inefficient decrease could have been due to the following factors:
  - Size of the evidence file
  - Version of operating system
  - Version of NSRL hash set (e.g., Minimal v. Modern)
  - Forensic program used for analysis (e.g., FTK®)
- Tables 11 and 17 displays how consistent the various tools are in comparison to each other
  - Magnet Axiom™ was omitted due to its percent decreases not meeting the 50% efficiency threshold.
  - Four of the six programs were within two standard deviations.
    - Four of the six programs perform within an acceptable range of consistency and reproducibility.

Part Two: Project VIC & Custom Hash Sets

- FTK®, Magnet Axiom™, and Griffeye Analyze DI successfully imported the most recent hash set version.
- Cellebrite Inspector imported a previous hash set version.
- Autopsy® and EnCase™ Forensic could not ingest any version of Project VIC.
- All programs, except Autopsy®, used the custom hash sets to successfully target and locate files of interest.

Part Three: Hashes & Mobile Devices

- Saving the images directly to the iPhone's Photo Library changed the hash values resulting in no match when comparing to the custom hash set using HashCalc.
- Uploading the images to iCloud, then syncing the iPhone resulted in no change to the images' hash values.
- Saving the images on the iFit tablet, Google Photos, and the Samsung Galaxy, Gallery, both devices resulted in no change to the images' hash values.
- The extraction used with the devices was unable to be ingested into Griffeye to use its artificial intelligence to match the altered images from the iPhone to the original versions.

References

Interpol.int. Digital Forensics. Accessed August 3, 2023. https://www.interpol.int/en/How-we-work/Innovation/Digital-forensics#:~:text=Digital%20forensics%20is%20a%20branch,crucial%20for%20law%20enforcement%20investingations.

Joseph DP, Viswanathan P. SDOT: Secure Hash, Semantic Keyword Extraction, and Dynamic Operator Pattern-Based Three-Tier Forensic Classification Framework. 2023; 11: 3291-3306. https://doi.org/10.1109/Access.2023.3234434.

Miller C. A survey of prosecutors and investigators using digital evidence: A starting point. Forensic Science International: Synergy. 2022. https://doi.org/10.1016/j.fsisyn.2022.100296.

Projectvic.org. Accessed August 8, 2023. https://www.projectvic.org/photo-dna.

Projectvic.org. Accessed June 30, 2023. https://www.projectvic.org/project-vic.

Rowe NC. Testing the National Software Reference Library. Digital Investigation. 2012; 9: S131-S138.doi.org/10.1016/J.DIIN.2012.05.009.

Socha G, Shah S. Data Validation A crucial step toward controlling and understanding your data. Bolch Judicial Institute at Duke Law. 2018; 102(3): 7-10. https://judicature.duke.edu/articles/data-validation-a-crucial-step-toward-controlling-and-understanding-your-data/.

SWGDE.org. Glossary. Accessed June 5, 2023. https://www.swgde.org/glossary.

Wilson-Kovacs D, Rappert B, Redfern L. Dirty Work? Policing Online Indecency in Digital Forensics. The British Journal of Criminology. 2022; 62 (1): 106-123. https://doi.org/10.1093/bjc/azab055.

Acknowledgments

The North Carolina State Crime Laboratory, specifically the Digital Evidence section, for the opportunity to learn and conduct this research, as well as for their guidance and dedication throughout the course of the project.