Deduplicated Disk Image Evidence Acquisition and Forensically-Sound Reconstruction

Xiaoyu Du, Paul Ledwith and Mark Scanlon
Forensics and Security Research Group
School of Computer Science
University College Dublin
Dublin, Ireland
Email: {xiaoyu.du, paul.ledwith}@ucdconnect.ie, mark.scanlon@ucd.ie

***Abstract*—The ever-growing backlog of digital evidence waiting for analysis has become a significant issue for law enforcement agencies throughout the world. This is due to an increase in the number of cases requiring digital forensic analysis coupled with the increasing volume of data to process per case. This has created a demand for a paradigm shift in the method that evidence is acquired, stored, and analyzed. The ultimate goal of the research presented in this paper is to revolutionize the current digital forensic process through the leveraging of centralized deduplicated acquisition and processing approach. Focusing on this first step in digital evidence processing, acquisition, a system is presented enabling deduplicated evidence acquisition with the capability of automated, forensically-sound complete disk image reconstruction. As the number of cases acquired by the proposed system increases, the more duplicate artifacts will be encountered, and the more efficient the processing of each new case will become. This results in a time saving for digital investigators, and provides a platform to enable non-expert evidence processing, alongside the benefits of reduced storage and bandwidth requirements.**

I. Introduction

The digital evidence backlog has been long outlined as a significant impeding digital forensic challenge [1], [2], [3]. The average backlog in digital forensic laboratories around the world was from 6 months to 1 year in 2009 [1]. In the UK, the most severe example saw one case being delayed by more than 21 months in 2015, and in 2016, the backlog had exceeded four years in the extreme case in Ireland [4].

The complicating factors resulting in the mounting digital forensic backlog include [5]: (i) the increasing number of cases involving digital investigation; (ii) the number of digital devices requiring analysis per case; (iii) the increasing storage volume of each device; (iv) the diversity of digital devices, storage formats, file systems, and physical data locations, e.g., Internet-of-Things devices, wearables, cloud storage, remote storage, peer-to-peer file synchronization services, etc. [6].

Data reduction techniques can aid in decreasing the volume of data to be analyzed. Data deduplication is a data reduction technique used to optimize data storage and is particularly efficient when common data is encountered. In 2016, Neuner et al. [7] applied data deduplication techniques for digital forensics. The authors reported that the storage requirement can be decreased by up to 78% in a real-world scenario.

The benefits of transitioning to a cloud-based digital forensic process model include: i) Always Up-to-Date Software Resources; ii) Pooled Hardware Resources; iii) Resource Management; iv) Flexible Location and Time [8]. Additionally, a significant cost can be saved by law enforcement through centralizing the processing of digital forensic evidence [4].

In 2009, Beebe [9] highlighted the need to automate the digital forensic process. Automation comes at a great expense and has had limited impact to date [10]. During the last decade, there has been some progress on investigation automation [11], [12], [13], but much remains to be done.

In a previous publication, the centralized data deduplicated framework has been discussed, including the aims and requirements for the system, analyzing of advantages over traditional approach (storage saving, bandwidth saving, more efficient processing, etc.) [4]. In this paper, the focus is on the implementation of the proposed system, proving the hypothesis of image reconstruction for a deduplicated acquisition method through experiment; testing and analysis result.

A. Contribution of this Work

The contribution of this work can be summarized as follows:

- Forensically-Sound Disk Image Reconstruction - Whenever required, a complete, verifiable disk image can be reconstructed from the deduplicated data store.
- Contributing to the Viability of Digital Forensics as a Service (DFaaS) - This system acquires digital evidence to a cloud-based system, together with associated metadata facilitating centralized analysis;
- Performance Evaluation of Deduplicated Acquisition - Test results prove that deduplication saves more storage space as more devices are acquired. This also results in improved transmission times for remote acquisitions;
- Code released open-source[1].

II. Related Work

A. Digital Evidence Processing

There are three main categories of activity in the process of digital forensics: acquisition, analysis, and presentation [14].

[1]https://github.com/XiaoyuDu/dedupinforsec

This research is focused on the acquisition and analysis process. Most digital investigations begin by creating a forensic image. There are two different categories of forensic imaging; one is raw acquisition, and the other involves forensic containers. `dd`, `dcfldd` and `dc3dd` are common tools for creating raw images. The latter two tools generate hashes of encountered data for forensic analysis. Common forensic container formats are Expert Witness Format (EWF)/E01 and Advance Forensics Format (AFF). EWF is used by the EnCase forensic suite and commonly The Sleuth Kit (TSK) supports AFF image files. The difference between raw and container files are i) raw images only include the data itself, whereas container files include the data from forensic devices and associated metadata; ii) raw images are fixed size, whereas containers can often be compressed.

B. Data Deduplication

The ability to store larger volumes of data in smaller physical disk space is a desirable behavior across numerous usage scenarios. The ever-increasing volume of storage devices is not expected to slow, and according to Kryder and Kim [15], it may continue to accelerate faster than expected. Data deduplication techniques have been widely used in a number of corporate products and systems mainly focusing on improving storage utilization.

Within the field of digital forensics, there is a significant amount of concern placed on having to store the massive amounts of data collected with each acquisition [5], [16], [9]. The increase in data also affects the time taken to both acquire and to analyze the data. If this problem continues to be left unaddressed, it may lead to case processing capacity problems in the future. Large data volumes also hinder, or entirely render infeasible, remote evidence acquisition. This leads to a requirement for techniques to reduce this amount of data within every acquisition.

C. Centralized Digital Evidence Processing

Centralized digital forensics is a relatively new approach that shows great potential to improve the efficiency of digital forensic investigations. The "Forensic Cloud" concept, a friendly work environment for investigators without special forensic tools knowledge, was proposed by Lee and Un in 2011 [17]. And in 2012, these same authors have implemented a cloud-based service for index search [18]. One implemented Digital Forensic as a Service (DFaaS) system is Xiraf (and its successor, HANSKEN), which has been built by the Netherlands Forensics Institute (NFI) [19], [20]. This system is implemented based on a model proposed by Kohn et al. in 2013 [21]. It facilitates the non-expert triage of forensic evidence while waiting for expert case analysis.

D. Automation and Intelligent Investigation

In 2009, Garfinkel developed an automated artifact extraction tool, Fiwalk [22]. It is implemented based on The Sleuth Kit[2] (TSK). The program executes TSK commands as a subprocess and processes the results. This project produces an XML output describing the contents of a forensic image file and enables efficient data analysis. An automated disk investigation toolkit (AUDIT) was created in 2014 [23]. This toolkit aids in automating part of the process, but leaves much to be achieved for full automation. To date, the influence of automatic tools to forensic investigations has been limited, as most evidence processing still requires human decision making. Applying automation to digital forensic investigation brings up many challenges [24].

E. Existing Deduplication Systems

Teleporter [25] is a remote evidence acquisition system. It operates on block-level deduplication, and the system is described as analytically-sound. Analytically-sound is a concept referring to forensically-sound individual artifact acquisition and analysis. The definition of a forensically-sound duplicate is a bit-for-bit copy of a drive, which contains every bit and sector of data on the physical level and does not alter any data.

The system presented in this paper goes one step further over Teleporter. Reconstruction tests have proven forensically-sound images can be recreated from a deduplicated evidence acquisition system, i.e., full disk hashes match *without* acquiring all artifacts from the suspect device. Even though the system is designed for forensically-sound acquisition, it can still operate for selective evidence acquisition, which targets potential pertinent data in the first instance.

Forensically-sound image reconstruction is necessary in a deduplicated remove digital evidence acquisition system. Because it proves that the acquisition includes all bytes of the original storage data. More importantly, the court-admissibility of digital evidence can be maintained.

III. Methodology

The process is designed to be user-friendly facilitating non-expert evidence acquisitions and can potentially be used by digital field triage, as outlined in [26].

A. Tools and Techniques

`Pytsk`[3] is a Python binding for The Sleuth Kit. The system outlined as part of this paper was developed using the `pytsk` library for recursively searching for files, their extraction and hashing them from the disk image. `Pytsk` supports various file systems including NTFS, FAT12, FAT16, FAT32, exFAT, Ext2, Ext3, HFS, etc.

B. Deduplicated Data Acquisition

In the proposed system, prior to acquisition, files and data are hashed and compared with a centralized, known-file database to eliminate common files. The data acquired from the evidence devices include the files, slack space (at the disk level and block level) and unallocated space. The data collection process only collects the unique artifacts encountered in the evidence, saving bandwidth and storage space. This acquisition process acquires more than just a copy of evidence

[2]https://www.sleuthkit.org

[3]https://github.com/py4n6/pytsk

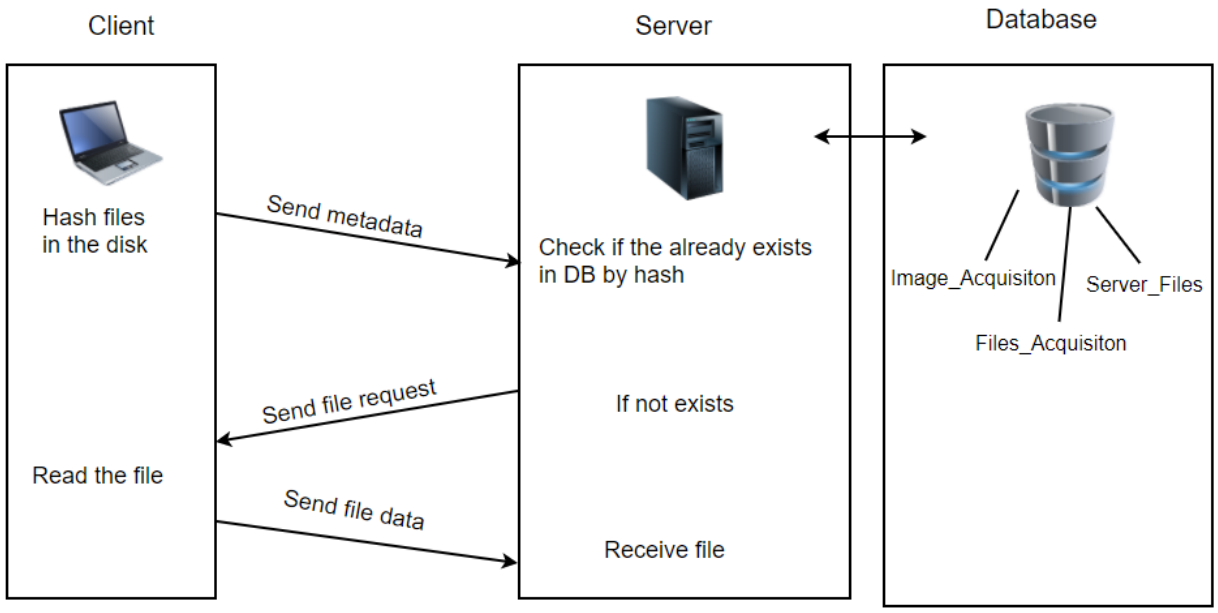

Fig. 1. Deduplicated Evidence Acquisition Process

artifacts, it has also completed part of the analysis, i.e., the calculation of the artifact hashes and the indexing of associated metadata for future examination. Figure 1 shows the process of evidence data transmitted from the client to the server. The steps include:

1) Metadata Extraction - For each artifact, its metadata together with the calculated hash is sent to the server. Hashes are used to compare against previously acquired artifacts, and the metadata is saved into the database.
2) Data Deduplication - If the artifact already exists on the server, the file data does not need to be reacquired. If it does not exist, then the server will send a data request to the client.
3) Data Transmission - For artifacts on the disk drive, the process of hashing the file data, checking with the server, and sending file data is multi-threaded.
4) Integrity Check - In case of data loss or corruption, every artifact is verified after it is acquired by the server. If the hash is not the same, then the server will send a request again.
5) Artifacts and Metadata Storage - Collected metadata and artifacts are saved on the server ready for reconstruction when needed. The metadata storage has three uses: i) it supports the deduplicated evidence acquisition; ii) metadata examination is necessary for digital forensic investigation; iii) it collates all the metadata for each acquisition together.

C. Disk Image Reconstruction

In comparison with the current state of art and alternative approaches, the approach outlined in this paper progresses one step further, i.e., forensically-sound complete disk image reconstruction. In the proposed system, data imaging is not achieved through an entire bit-for-bit copy, but it does afford the same forensically-sound disk image to the investigator. The acquisition result can be verified by comparing the hash of recreated disk image to the original one. For the purpose of disk image reconstruction, there are three constituents of binary data necessary. The file data, the block-level slack space, and the unallocated space on the disk. This reconstructed image can subsequently be verified against the original drive by comparing their hash values.

Figure 2 shows the process for image reconstruction. The forensic artifacts from each acquisition and the metadata stored in the database are necessary for recreating a forensically-sound image. To recreate an image, the system first needs a specific *acquisition id* and based on this *id*, the information such as image size and data storage locations can be retrieved. A blank staging image is first created and subsequently each of the artifacts are placed at the same specific physical block offset as in the original disk. Finally, a hash is generated and compared with the original device's to verify successful reconstruction.

D. Log Files

Each of the operations on the server generates auditable log files for analyzing the performance and verifying the accuracy of the system. The acquisition log records the data verification, collection time, transmission speed, duplication ratio, etc. The reconstruction log includes the reconstruction time, result, hash of the complete disk image, etc.

E. Benefits of this Approach

- Centralized Digital Evidence Processing - Digital artifacts analysis results are stored on the centralized server. As a

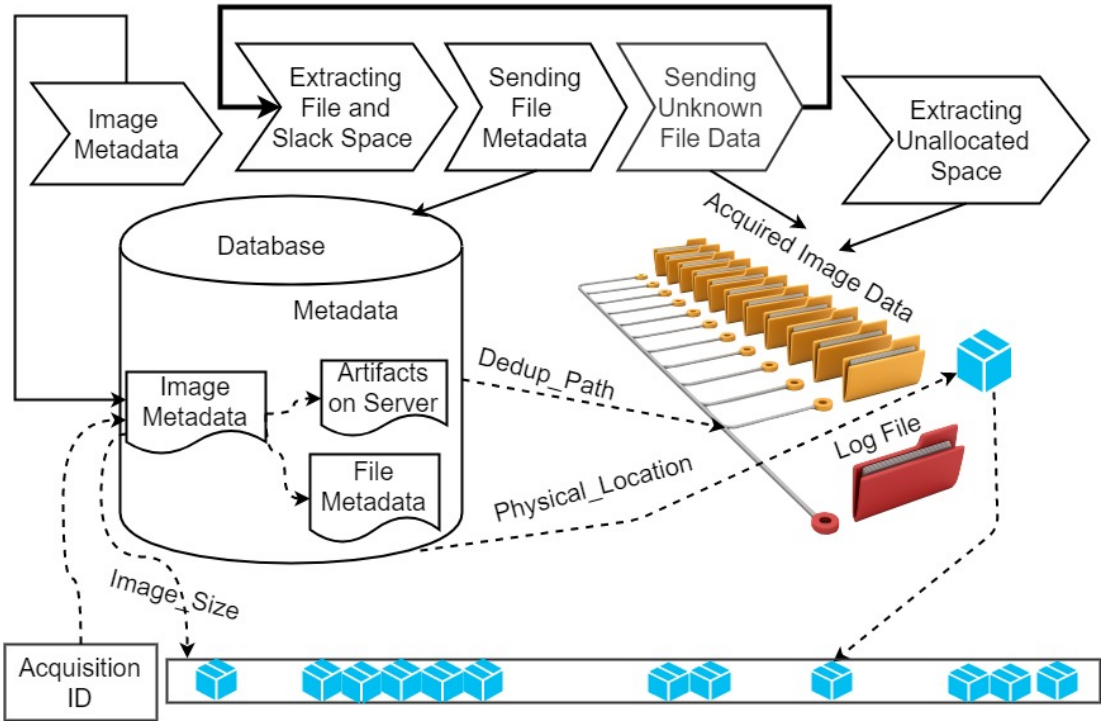

Fig. 2. Image Reconstruction from Deduplicated System

result, they each artifact need only to be processed once;

- A Model for Non-expert Acquisition and Analysis - This system processes digital evidence automatically;
- On-the-Fly Incriminating File Detection - Known illegal artifacts can be detected *during* the acquisition step, rather than after complete acquisition;
- The Bigger, The Better - The more acquisitions performed using the proposed system, the higher the expected duplication rate encountered and the faster future acquisitions will become.
- Intelligent Digital Evidence Analysis - Stored expert analysis and previous data analysis can be used to train artificial intelligent activity/event patterns to detect suspicious file artifacts automatically.

IV. Results

A. Prototype Setup and Test Data

The prototype system is running on a server with Ubuntu 16.04.2 LTS operating system, Linux 4.4.0-121-generic kernel, x86-64 architecture.

For testing the performance of evidence acquisition and reconstruction, 12 images were created with various file systems. The images were created through *dd* copying data from a USB drive. Various duplication ratios were created to test the performance of duplicated data. The test data information is shown in Table I.

TABLE I
Test Data

Images	Size	No. of Files	File System
A_Image	150MB	477/491/513	FAT
D_Image	2/8/16GB	36/234/244k	NTFS
Windows_PE	2GB	1645	NTFS
Windows7_Image	10GB	48k/49k/50k/58k	NTFS
Windows8_Image	10GB	81k	NTFS

B. Acquisition Speed

Through iterative acquisitions of the images, measurements of the acquisition speed and time were recorded, alongside the acquisition date and time, and the encountered duplication ratio. Figure 3 shows the average acquisition speed for each image. Each of the created images were acquired several times. Analysis of the results identified three different factors influencing the speed:

1) The Image - The overall size of the image, the number of files on the image, and the ratio of small files compared with large files;
2) Duplication Ratio - This determines how much data has to be transmitted to the server;
3) Execution Environment - Network bandwidth, client computer processing speed, storage hardware performance, etc.

Figure 4 shows a comparison of two speeds; one is actual hardware read-speed, the other is the effective speed, i.e., the throughput of the system factoring in the speed enhancements provided through deduplication. This effective speed is consistently faster than the actual speed. The formulas below demonstrate precisely what constitutes these two speeds (the reduced size is the original data less the duplicated data which

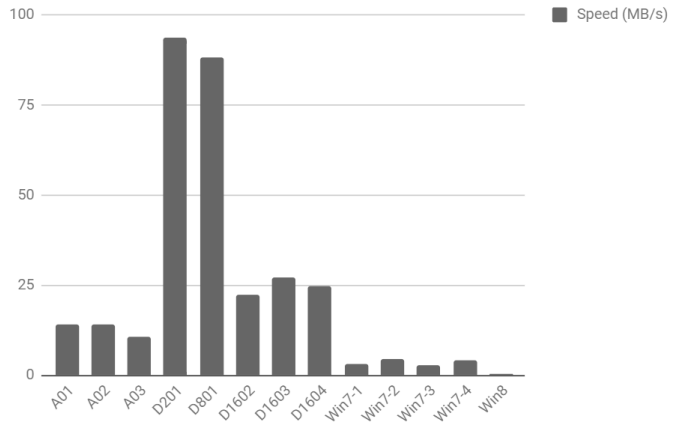

Fig. 3. Evidence Acquisition Speed of Each Image

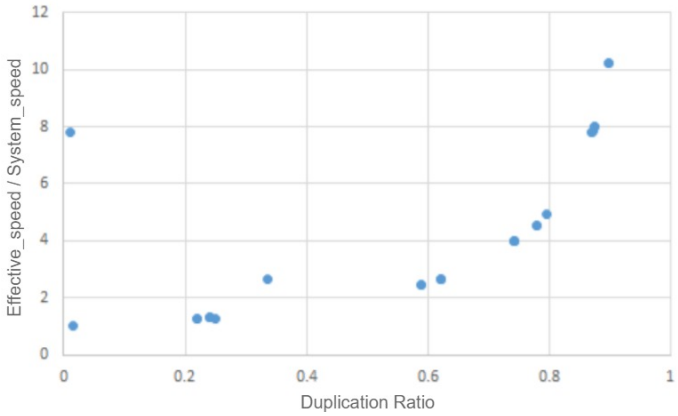

Fig. 5. Duplication Ratios and their Impact on Speed

does not need be re-uploaded).

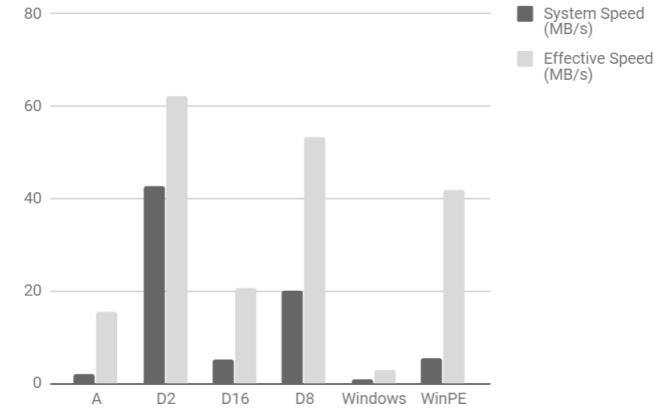

Fig. 4. System Speed and Efficient Speed Comparison of Each Image

Data deduplication improves the system speed significantly. Figure 5 illustrates that the higher the duplication ratio, the faster the effective acquisition speed. Test results show the speed can be ten times disk-read speed when the duplication ratio is approximately 90%.

When interpreting the above acquisition speeds, it is important to note that some preprocessing of the data has already taken place in addition to the acquisition. For example, the file system metadata for each artifact has already been extracted and recorded in the database including its path, filesize, hash, block locations, etc. It is also possible to extend the current proof-of-concept system to highlight known illegal files *during* the acquisition phase of the investigation, flagging pertinent information to the investigator at the earliest stage possible.

C. Storage Saving

Data deduplication not only speeds up data transmission but also saves system storage space requirements. The system is designed for big volume data storage. As the volume of data collected grows, the more duplicates are encountered, the more storage is saved. In this test, the first acquisition of all the images, the system had to take 20% extra storage space; while when 1TB digital evidence acquired, this system saves up to 32% storage space.

D. Image Reconstruction

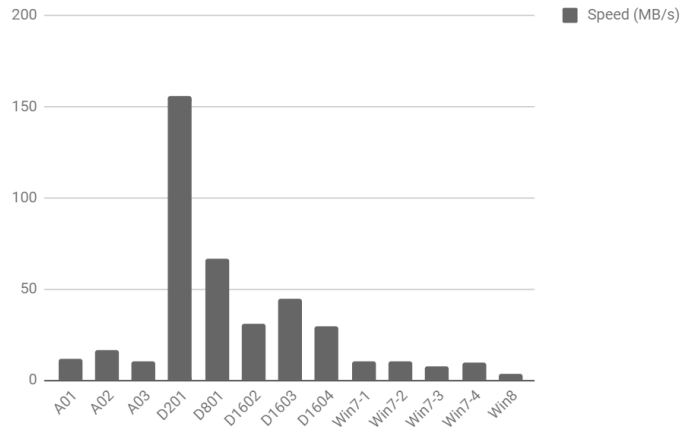

Fig. 6. Reconstruction Speed of Each Image

Hundreds of validated disk image reconstructions from the deduplicated data store have been successfully performed. Figure 6 shows the average speed of each disk image reconstruction. The speed varies due to the aforementioned influencing factors. The fastest individual reconstruction attempt during testing was over 150MB/s. The average of all the reconstruction speed is 43.78MB/s, which can be improved upon in the future through the employment of RAID storage enabling faster disk I/O. Disk image reconstruction may only be necessary if incriminating evidence is discovered. The Windows Preinstallation Environment (PE) image used is faster than Windows image as the number of small files is significantly less. D_Image was faster than each of the others with average reconstruction speed 113.67 MB/s due to it containing a higher proportion of large files.

V. Conclusion

This research explores a novel approach to collect digital evidence from a variety of devices and evaluates the storage, speed through several forensically-sound evidence acquisitions. The factors that influence the performance of the proposed system were also evaluated. From the analysis of the results, the summary is as follows: i) Acquiring data from suspect devices is complete and verifiably accurate; ii) Forensically-sound complete disk image reconstruction was achieved for all test data; iii) As a byproduct of the deduplicated acquisition process, evidence preprocessing has also taken place including metadata extraction and artifact hashing; iv) The performance is better for disk images containing a high proportion of large files; for complete operating system images, the speed achieved shows great promise for the technique, but still needs to be improved to be viable. In a remote acquisition scenario, i.e., acquiring a forensic image over the Internet, the acquisition speed is reasonable when compared with typical broadband upload speeds.

A. Future Work

The performance of evidence acquisition can be improved in future research. This can be achieved through the employment of improved hardware infrastructure and the introduction of a level of deterministic risk. The latter would involve the use of partial artifact hashing for deduplication purposes – greatly expediting the overall acquisition time. While this risk is introduced at the acquisition step, full-disk hashing can be used to verify forensically-sound reconstruction. Additionally, a local client-side data store on the acquisition client can be implemented to check the existence of duplicates resulting in minimizing the network traffic. This system can also benefit from integrating evidence prioritization targeting potentially pertinent evidence at the earliest stage of acquisition.

While some effort has been made towards the automation of the analysis step of a typical investigation, significant expert human analysis is still required. Recording the expert digital investigators' categorizations, decisions, and analyses in the centralized database enables a number of interesting future research directions. For example, leveraging these expert decisions to train artificial intelligence and big data analytics-based techniques to enable automated evidence processing.

References

[1] E. Casey, M. Ferraro, and L. Nguyen, "Investigation Delayed Is Justice Denied: Proposals for Expediting Forensic Examinations of Digital Evidence," *Journal of forensic sciences*, vol. 54, no. 6, pp. 1353–1364, 2009.

[2] J. I. James and P. Gladyshev, "Challenges with Automation in Digital Forensic Investigations," *arXiv preprint arXiv:1303.4498*, 2013.

[3] A. Shaw and A. Browne, "A Practical and Robust Approach to Coping with Large Volumes of Data Submitted for Digital Forensic Examination," *Digital Investigation*, vol. 10, no. 2, pp. 116–128, 2013.

[4] M. Scanlon, "Battling the Digital Forensic Backlog through Data Deduplication," in *Proceedings of the 6th IEEE International Conference on Innovative Computing Technologies (INTECH 2016)*. Dublin, Ireland: IEEE, 08 2016.

[5] D. Lillis, B. Becker, T. O'Sullivan, and M. Scanlon, "Current Challenges and Future Research Areas for Digital Forensic Investigation," in *The 11th ADFSL Conference on Digital Forensics, Security and Law (CDFSL 2016)*. Daytona Beach, FL, USA: ADFSL, 05 2016, pp. 9–20.

[6] M. Scanlon, J. Farina, and M.-T. Kechadi, "Network investigation methodology for bittorrent sync: A peer-to-peer based file synchronisation service," *Computers & Security*, vol. 54, pp. 27 – 43, 10 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016740481500067X

[7] S. Neuner, M. Schmiedecker, and E. Weippl, "Effectiveness of file-based deduplication in digital forensics," *Security and Communication Networks*, vol. 9, no. 15, pp. 2876–2885, 2016.

[8] X. Du, N.-A. Le-Khac, and M. Scanlon, "Evaluation of Digital Forensic Process Models with Respect to Digital Forensics as a Service," in *Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS 2017)*. Dublin, Ireland: ACPI, 06 2017, pp. 573–581.

[9] N. Beebe, "Digital forensic research: The good, the bad and the unaddressed," in *IFIP International Conference on Digital Forensics*. Springer, 2009, pp. 17–36.

[10] S. L. Garfinkel, "Digital forensics research: The next 10 years," *digital investigation*, vol. 7, pp. S64–S73, 2010.

[11] C. Hargreaves and J. Patterson, "An automated timeline reconstruction approach for digital forensic investigations," *Digital Investigation*, vol. 9, pp. S69–S79, 2012.

[12] R. In de Braekt, N.-A. Le-Khac, J. Farina, M. Scanlon, and M.-T. Kechadi, "Increasing Digital Investigator Availability through Efficient Workflow Management and Automation," Little Rock, AR, USA, pp. 68–73, 04 2016.

[13] T. Thornbury, M. A. Brock, J. D. Redmon, and J. W. Texada, "Automation of collection of forensic evidence," Jun. 13 2017, uS Patent 9,680,844.

[14] C. Altheide and H. Carvey, *Digital forensics with open source tools*. Elsevier, 2011.

[15] M. H. Kryder and C. S. Kim, "After hard driveswhat comes next?" *IEEE Transactions on Magnetics*, vol. 45, no. 10, pp. 3406–3413, 2009.

[16] S. Neuner, M. Mulazzani, S. Schrittwieser, and E. Weippl, "Gradually improving the forensic process," in *Availability, Reliability and Security (ARES), 2015 10th International Conference on*. IEEE, 2015, pp. 404–410.

[17] J. Lee and D. Hong, "Pervasive forensic analysis based on mobile cloud computing," in *Multimedia Information Networking and Security (MINES), 2011 Third International Conference on*. IEEE, 2011, pp. 572–576.

[18] J. Lee and S. Un, "Digital forensics as a service: A case study of forensic indexed search," in *ICT Convergence (ICTC), 2012 International Conference on*. IEEE, 2012, pp. 499–503.

[19] R. Van Baar, H. van Beek, and E. van Eijk, "Digital Forensics as a Service: A Game Changer," *Digital Investigation*, vol. 11, pp. S54–S62, 2014.

[20] H. van Beek, E. van Eijk, R. van Baar, M. Ugen, J. Bodde, and A. Siemelink, "Digital Forensics as a Service: Game on," *Digital Investigation*, vol. 15, pp. 20–38, 2015.

[21] M. D. Kohn, M. M. Eloff, and J. H. Eloff, "Integrated digital forensic process model," *Computers & Security*, vol. 38, pp. 103–115, 2013.

[22] S. Garfinkel, "Automating disk forensic processing with SleuthKit, XML and Python," in *Systematic Approaches to Digital Forensic Engineering, 2009. SADFE'09. Fourth International IEEE Workshop on*. IEEE, 2009, pp. 73–84.

[23] U. Karabiyik and S. Aggarwal, "Audit: Automated disk investigation toolkit," *The Journal of Digital Forensics, Security and Law: JDFSL*, vol. 9, no. 2, p. 129, 2014.

[24] J. I. James and P. Gladyshev, "Automated Inference of Past Action Instances in Digital Investigations," *International Journal of Information Security*, vol. 14, no. 3, pp. 249–261, 2015.

[25] K. Watkins, M. McWhorte, J. Long, and B. Hill, "Teleporter: An Analytically and Forensically Sound Duplicate Transfer System," *Digital Investigation*, vol. 6, pp. S43–S47, 2009.

[26] B. Hitchcock, N.-A. Le-Khac, and M. Scanlon, "Tiered forensic methodology model for digital field triage by non-digital evidence specialists," *Digital Investigation*, vol. 16, no. S1, pp. 75–85, 03 2016, proceedings of the Third Annual DFRWS Europe.