Exploring Solutions Put Forth to Solve Computer Forensic Investigations of Large Storage Media

A.Z. Tabona and W.B. Glisson

School of Humanities
University of Glasgow
e-mail: 1005599z@student.gla.ac.uk, brad.glisson@glasgow.ac.uk

Abstract

The capacity of digital storage media is growing at a phenomenal rate, leading to an increase in the overall time it takes to process a typical digital forensics investigation. Conventional tools and techniques simply do not cater for the size of potential evidence that investigators have to analyse. With digital evidence being available on an increasing number of digital media types, ranging from portable media players, to Global Positioning System (GPS) devices, to rack-mountable servers, in addition to the fact that there is a rising trend for digitizing information in the business world, the problem is only getting worse. This paper endeavours to initiate an investigation into the current solutions put forth to solve computer forensic investigations of large storage media for the purpose of stimulating ideas and encouraging expansion of current solutions within the research community.

Keywords

Forensics, Analytics, Datasets, Triage, Search, Distributed Processing, Hashing, Large-Scale, Data Mining

1. Introduction

Storage media vendors are answering calls from the market to have more compact devices that give customers increased storage for their dollar value. At the time of writing, a 3TB (Terabytes) hard disk drive costs £140 ($220) (dabs.com, 2011). The past few years have seen a decrease in the physical size of storage media but an increase in the storage capacity at the same time. Nowadays, it is not uncommon to find pervasive media devices that hold between 16GB to 160GB (Gigabytes) worth of data (Apple, 2011), or entry-level desktop computers that come pre-built with at least 250GB hard disk drives (Dell, 2011). This increase in storage media has a direct impact on the proliferation of digital evidence in a typical computer forensic investigation. Statistics from the FBI (FBI, 2010a) show that the number of Terabytes they processed for the 2009 Financial Year doubled from two years prior to reach a staggering 2334 Terabytes worth of data. By way of comparison, think of the entire contents of an academic research library (each and every book, journal, transcript, etc.) being the equivalent of 2 Terabytes. The notion of data doubling every two years is further corroborated by Roussev (2009).

The traditional approach to a computer forensics investigation (seize, image, search) is no longer viable for large-scale examinations (Rogers, 2006; Rowlingson, 2004). Storage media capacity has increased at such a rate that computer forensic investigators are unable to keep up. Their productivity throughput and ability to maintain a reasonable backlog have also suffered as a result of storage media being too large for manual analysis (FBI, 2010a). In addition, a typical forensic investigation is likely to encompass multiple devices or machines, ranging from smartphones to laptops to large-scale financial database servers. Gone are the days when a digital forensics team had the resource to analyze each and every piece of storage media looking for potential evidence.

Investigators might agree that current implementations of industry tools (such as AccessData's FTK and Guidance Software's EnCase) have failed to implement any substantial time saving techniques. Their inability to sufficiently handle terabyte datasets leaves investigators in a situation where even preliminary forensic tasks take an inordinate amount of time. Researchers have proposed a number of solutions to the problem, including hash lists of known files (Roussev, 2009), triage tools and models (AccessData, 2011; FBI, 2000; IDEAL Technology Corp., 2011, Rogers, 2006; Microsoft, 2011; SPEKTOR Forensic, 2011), de-centralized parallel processing (van den Hengel, 2008; Richard and Roussev, 2006; Liebrock, 2007; Roussev et al., 2009), data mining (Beebe and Clark, 2005; Rao, 2010; Abraham, 2006), data sampling (Mora and Kloet, 2010), data analytics and traditional search, filter and categorization techniques (Tanner and Dampier, 2010; Beebe and Clark, 2005; Beebe and Dietrich, 2007; Beebe, 2009; Pollitt, 2010), implementing layers of abstraction within the system (Carrier, 2003) and a methodology that promotes the sharing and re-use of knowledge gained from past forensic examinations (Tanner and Dampier, 2010). As well as the technical challenges faced by investigators when dealing with the increasing capacity of storage media, there are also the legal challenges to contend with (Spafford, 2006; Sommer, 2004). Investigators are under increasing pressure from the justice system to provide evidence quickly, when they do not understand the scale of the data that investigators need to analyse. In addition, the economics of handling evidence are also a challenge; the costs associated with printing terabytes worth of data as well as managing and maintaining the evidence itself (Sommer, 2004). There is a dire need for solutions that assist digital forensic investigators in reducing the dataset that they analyse by removing 'noisy' data and helping them to quickly visualize only data of relevance. This notion is further corroborated by Riley et al. (2008) who demonstrate how traditional preliminary forensic analysis tasks such as imaging eat into investigation time. Experiments by these authors show that imaging a 250GB hard drive takes in excess of 1.5 hours.

This paper is organized as follows. Section 2 looks at hashing techniques and Section 3 discusses triage. Section 4 discusses the different methods that involve the retrieval and analysis of data, while Section 5 contains a discussion about distributed forensics and the importance of parallel processing. Section 6 highlights the way things stand now and future work, whilst exploring a holistic design that allows various solutions to work as a cohesive unit. Finally, the author's conclusions are documented in Section 7.

2. Hashing

Hashing is a technique used for data integrity and verification matching of known data. The way hashing works it to take a random string of binary data as input and generate a unique alphanumeric value (known as the message digest) as output, typically in a SHA1 or MD5 format (Carrier, 2005). Hashing is uniformly applied in the computer forensics world to hard drive volumes or individual files for data integrity purposes or to compare files to determine if a dataset contains "known" objects (Carrier, 2003; Roussev, 2009). The concept is to compare the hashes of every file found on a hard disk with a database of hashes to see which ones match with a hash that is known to be "good" or "bad". Such hash databases would be pre-compiled and verified manually.

The National Software Reference Library (NSRL) (National Software Ref. Library, 2011), maintained by the US National Institute of Standard and Technology (NIST), is a comprehensive list of hashes belonging to common operating systems and applications as well as software that may be considered malicious. Unlike the US National Drug Intelligence Center's HashKeeper database (USDOJ, 2011), the NSRL does not contain a list of illicit data, such as known child pornography images. Both these databases are used extensively by investigators around the world to filter out 'noisy' data that does not need to be examined because it can quickly be determined if that data is considered to be good or bad. In addition to NSRL and HashKeeper, vendors of commercial forensics tools also maintain their own versions of hash databases that contain a list of known hashes for other software.

Hashing forms the basis of many commercial and open source forensic software packages. AccessData's FTK software, for example, has what they call the KFF (Known File Filter) which uses the NSRL and HashKeeper databases, allows you to import your own set of file hashes and customize the alerting options for different types of hashes (AccessData Forensic Toolkit. 2010). There are typically two approaches to this data reduction technique; 'known objects' which performs a comparison to determine if there is a direct match and, 'similar objects' which compares whether file 'A' is similar to any other file within the given dataset. These approaches are usually called 'simple hashing' and 'fuzzy hashing'.

2.1. Simple Hashing

Simple hashing involves comparing the hashes of known files, such as those found in the NSRL (National Software Ref. Library, 2011) and HashKeeper (USDOJ, 2011), with those found within the dataset. If a set of file hashes match then that data is 'removed' from the dataset that is displayed to the investigator. Simple hashing also allows file fragments to be found by splitting the file into multiple blocks and keeping a list of the hashes of each individual block to compare with data fragments found in unallocated space, for example (Roussev, 2009). The disadvantage of hashing is that, since the hash list is processed in memory in a sequential manner, the larger the dataset grows the worse the query performance throughput will be. While hash databases such as NSRL and HashKeeper aim to facilitate the fast processing of

the hash list by already coming as a sorted set, this does little to help in the long term as the hash database grows.

Roussev (2009) suggests using Bloom Filters to counteract this problem which operate by using a set of independent hash functions to generate a value for elements in a vector of size *m* based on a given input string. By using each hash function to generate a hash on a specific element, a comparison can then be made to see whether the element has its bits set to 1 or 0 when a lookup is made. Bloom Filters allow for fast processing of hash lists, but may generate a minimal rate of false positives. Roussev (2009) gives an example of increasing the number of hashes in a set from 50 million to 500 million and only having a false positive rate of 0.2 per cent. While simple hashing techniques allow known files to be filtered out at a faster rate than a manual analysis, they do not assist an investigator in finding files of a similar type.

2.2. Fuzzy Hashing

Fuzzy hashing allows two datasets to be compared for similarity. Characteristics are drawn from each object in the dataset and then compared for likeliness. The way it operates is to split each file into multiple chunks and generate a corresponding 6-bit hash for each chunk which are then concatenated into one larger base64 encoded hash. Roussev (2009) points out that fuzzy hashing techniques are successful in finding small objects and similar versions of files, but they suffer when it comes to finding similarities in larger files.

Related to fuzzy hashing is the concept of data fingerprinting, which generates a signature of file data that is more tolerant to changes in the original file (Roussev, 2009). This is different to the design of hashing which generates a different output if even one bit of the input is changed. Fingerprinting operates on the same lines as fuzzy hashing whereby characteristic features are selected for each object and then collated and compared for similarity at a binary level. This is highly beneficial in determining if a file has changed format. For example, consider a static XML file that is viewed in a web browser and then saved as a TXT file. Even though the tags will be removed as part of the conversion, the text should remain the same and produce the same set of characteristics as the original static XML file, allowing the versions to be correlated.

Lejsek et al. (2010) argue that hash lists are unsuitable for the purposes of video identification due to the lack of resources available to build and maintain a hash database of video content and that since there are so many different variations of such content, hashing video files would simply not be scalable in the long term. The authors propose a video identification solution whereby the video file is split into different points, each of which are encoded to have a unique fingerprint which can then be compared with a database of known video fingerprints. This automated process would save the investigator having to watch each and every video file to determine its category; it would essentially allow them to pinpoint video files that were of a pornographic nature for example. Lejsek et al.'s experiments show that in a sample dataset containing over 25,000 hours of video content, they were able to

classify the files 70% faster than a manual analysis. This is advantageous in allowing the investigator to get the job done in less time while saving on productivity costs. The disadvantage of such a technique is that as the size of the collection increases, the quality of the results decreases, therefore requiring constant fine tuning by the investigator.

3. Triage

Adapting the Oxford English dictionary's (2011) definition of 'triage' to the field of computer forensics, it could be interpreted as: 'A method whereby items are ranked in order of priority or importance in order to determine which should be processed first'. In investigations that involve kidnap or murder, time is of the essence as it could mean the difference between life and death. Rogers et al. (2006) propose The Cyber Forensics Field Triage Process Model (CFFTPM), a formalized methodology that was built upon years of real-world experience where various approaches have been tried and tested in the field. The authors demonstrate how their model can be used to collect various evidentiary materials and be used in several specific cases including child pornography, drug, and fraud based investigations. A benefit of this methodology is that it allows for a feedback loop between first and second line investigators. Information can be fed back to each other, allowing keyword and file type searches to be fine-tuned based on new information that is gathered from the initial triage process.

Industry tools such as AccessData's AD Triage (AccessData, 2011), the FBI's ACES (Automated Computer Examination System) (FBI, 2000) and Microsoft's COFEE framework (Microsoft, 2011), as well as triage devices like SPEKTOR (SPEKTOR Forensic, 2011) and STRIKE (IDEAL Technology Corp., 2011), allow first responders to collect relevant data quickly and in a forensically sound manner. Such tools can however be considered immature and will require time to evolve, based on user experience and technology advancements.

3.1. Image Recognition

Choudhury et al. (2008) propose a novel skin tone detection algorithm that analyses images to determine if they could be of a pornographic nature. The results of their experiments are encouraging and show accuracy rates of over 78%. Such a technique would be useful for first responders in a child pornography case where they could quickly run a 'triage' scan using a tool that implements this algorithm to determine which storage media is likely to contain illicit images. The investigator would then only have to seize, image and analyse those drives.

Chen et al. (2005) offer a Content Based Image Recognition (CBIR) technique that extracts properties from previously identified illicit images and uses these properties when searching target drives for contraband. The process will identify those images that share the same properties as other contraband and that may have had their orientation, quality, or size properties altered. This is useful in a triage scenario when

the investigator needs to quickly determine if a storage media contains illegal images.

3.2. Selective Data Acquisition

Lee et al. (2009) propose a methodology that promotes selective data acquisition, alleviating the need to duplicate or image the entire storage media and allowing for the investigation of relevant data only. Indeed, the process of selective data acquisition is considered to be one of the solutions to the challenges investigators face of dealing with significant amounts of data (Beebe, 2009). The Phased Investigation Methodology (PIM) is broken down into four steps; target selection and pre-investigation, tracing recent computer usage, computer usage pattern analysis, investigation of user-based file contents. At each step, the investigator can analyse only the data that is required and decide whether to move onto the next step or if all or parts of the system should be excluded from analysis altogether.

The drawback of using triage is the potential of missing some exculpatory evidence during the preliminary stages of deciding which machines or devices should be examined first. Additionally, forensic triage tools often come pre-configured with 'default' options or require pre-investigation customization. Opening up tools to investigator customization is beneficial in terms of allowing them to define what data is to be collected, but at the same time generates a risk of investigator bias.

4. Data Analytics

4.1. Scoped Search Methods

The feature set of the current generation of forensic tools already offer some time saving techniques for dealing with large data sets that are built around search mechanisms. File carving, for example, is a process whereby the tool automatically carves out files of interest based on their starting and ending file signature (Richard and Roussev, 2006). This alleviates the investigator from needing to manually sift through large amounts of hexadecimal values looking for obfuscated data. The disadvantage of such a process is that it may end up carving out duplicate files; thumbnails from larger JPEG files, images within PDF documents, etc. Consequently, with carving there is a tradeoff between looking at certain files twice versus missing a possibly crucial piece of evidence that is hidden amongst unallocated space.

File categorization techniques allow the investigator to focus on particular pieces of evidence. Sorting, filtering and file categories are often used in current forensic toolsets to search for particular information and reduce the dataset being displayed to the investigator. Sorting allows the investigator to group files by various fields, including file type, size, hash value, path, etc. Filtering allows only data that meets certain criteria to be displayed and file categorization displays files in a hierarchical format based on their category (e.g. Documents, Images, Deleted Items, etc.).

Searching allows the investigator to look for evidence containing a specified keyword or set of keywords using the 'AND' or 'OR' logical operators.

There are typically two types of search; what is known as 'live' search where the entire dataset is searched in real-time and 'indexed' search where the dataset, having previously been indexed, is searched for items that correspond to a given keyword. Garfinkel (2010) refers to these techniques as the "Visibility, Filter and Report" model and argues that while the investigator is able to search for specific items, they are not able to quickly prioritize the data that they find. Parallelizing such an approach is also not possible with current tools (Garfinkel, 2010), meaning that the problem is going to get worst as the capacity of storage media continues to increase. Furthermore, encrypted data, compressed files and embedded text formats cannot be evaluated using search utilities because they operate based on a string of plain text. Regular expressions can help with this by offering a method of finding text that matches a given data pattern. For example, the following regular expression will find all IP addresses in a search: \b\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}\b.

A problem associated with the use of a search utility is that it is assumed that the computer forensic investigator knows explicitly what keywords and phrases to use in the search. In addition, the time it takes for them to manually review these results is also a problem (Beebe, 2009). Relying on the human factor alone for each individual case would be unwise. This is where having pre-defined keyword lists based on the type of case being investigated would be beneficial, in addition of course to the investigator's intuition and an automated collection of keywords derived from the evidence itself.

Being dependent on conventional search methods will only make the problem worse in the long term, as datasets continue to grow at astounding rates. One approach to the problem of analysing large datasets is to improve the efficiency of the analytical approaches used today (Beebe, 2009; Pollitt, 2010). Beebe (2009) discusses how the current analytical approaches result in "underutilization of available computational power and high information retrieval overhead". The present-day implementations of search, retrieval and analytic algorithms are underutilized and do not scale to the full potential of today's computing platforms. It would be wise to borrow a leaf out of the information retrieval group's expertise and utilize an advanced algorithm for intelligent search mechanisms in digital forensics applications. This would allow relevant data to be presented to investigators more quickly whilst reducing the "noisy" results.

Text string searches occur at the physical level of the storage media, meaning the specified search string can be found in locations that are independent of the logical data structures, partitioned volumes, and data allocation flags (Beebe and Dietrich, 2007). This poses as another problem for conventional text based search tools in that they return an extremely high number of 'hits' (Beebe and Clark, 2007; Beebe and Dietrich, 2007), a lot of which bear no relevance whatsoever to the case in question. To solve this problem, there is a need for the search tool to either reduce the number

of irrelevant search hits returned or present the results in a way that allows the investigator to locate relevant hits quickly (Beebe and Clark, 2007).

Beebe and Clark (2007) propose an approach that builds on the latter and allows the investigator to locate more relevant hits more quickly by thematically clustering digital forensic text string search results. The algorithm would be responsible for automatically grouping false positives that are generated as part of a generalized keyword search, allowing the investigator to skip these false positives as they analyse the results. Beebe and Dietrich (2007) propose an approach which introduces additional system state transitions and operators. Additional levels of computer information processing (CIP) operators gather, classify, index and rank the hits before they are presented to a human for information processing. This is beneficial since less relevant evidence will be missed and can scale to large datasets. The disadvantage of such a technique is that there is a risk that the investigator will become over reliant on the computer interpreted information and bias towards certain hits.

Tanner and Dampier (2010) suggest a socio-technical approach to the problem of investigating evidence on large storage media. This approach involves the use of concept maps to generate and represent the relationships between information and expert knowledge. They argue that this knowledge can then be shared with other investigators and re-used in future examinations for selecting better search terms. The pros to such a solution are that it will help to narrow down the results that need to be analysed and offer a visual representation of them. The cons to this approach are that it is initially time consuming, requires a comprehensive learning curve and may result in duplicating work processes.

4.2. Random Data Sampling

To help reduce backlog produced as a result of investigators having to analyse large storage media, random data sampling can be used to calculate the number of items that the investigator needs to review from a given dataset. This will allow the investigator to have x% confidence that between n% and n% of the items on the storage media are contraband. The formula used to calculate the sample size that the investigator is required to analyse is shown below (Mora and Kloet, 2010):

$$n = \frac{N}{1 + N(e)^2}$$

n is the sample size, *N* is the population size, and *e* is the level of precision required. The formula above assumes a degree of variability of 0.5 and a confidence level of 95%. As an example, consider a dataset containing 2000 items (this is the population size) where the investigator wanted to know how many items he or she would need to review to be 95% confident that the remainder of the dataset contained contraband. Using the formula above, the sample size would be 333, meaning the investigator would only need to review this amount of items from the population as opposed to the entire population itself (Mora and Kloet, 2010).

The lack of random data sampling techniques in commercial forensic applications like Guidance Software's EnCase and AccessData's FTK, are putting the investigator at a disadvantage by requiring them to analyse a large set of data when, based on statistical theorems, they should really be able to only analyse a sample of that data to give them confidence that there is a specific type of material present on the evidence drive. Mora and Kloet (2010) propose using random data sampling as a triage technique whereby the first responder would review a given sample size from a data collection to determine whether that storage device or computer needed to be seized and taken back to the lab for further analysis.

4.3. Data Mining

Data mining is the process of finding and retrieving data from large datasets and presenting that data to the user in a useful and comprehendible format. Data mining can be sub-divided into three classes; (1) descriptive data modelling, (2) predictive data modelling, and (3) content retrieval (Beebe and Clark, 2005).

Descriptive data modelling involves the summarization and comparison of data with the aim of aggregating it into a smaller subset of data. Summarization is unsuitable for digital forensic purposes due to the risk of data loss during the aggregation stage. Comparison is more suited however, since it only involves determining the differences between two datasets. Predictive data modelling involves identifying specific data characteristics with the aim of anticipating future observations, and content retrieval data mining involves the retrieval of unstructured or semi-structured data sets (Beebe and Clark, 2005).

Data mining techniques for multimedia are of the utmost importance for digital forensics. Child pornography cases can involve thousands upon thousands of illicit images (FBI, 2010b), which is often too much of a burden for the investigator(s) to analyse manually. By using data mining, images of interest can be classified by type (human, building, car, etc.) and retrieved and displayed to the investigator accordingly.

One of the key benefits of data mining is the possibility to rank documents based on relevance to the investigation and prioritize the search 'hits' that the investigator will analyse. In addition, by using data mining techniques, costs and system and human processing time can be reduced and data analysis quality improved. Enhanced utilization of computing power and the ability to discover trends within a dataset that are normally hidden to humans, are also benefits (Beebe and Clark, 2005). The drawbacks of data mining potentially include advanced training to know how to mine properly and proper interpretation of the data. Since data mining converts data at a higher abstraction level error calculations are possible (Beebe and Clark, 2005).

Abraham (2006) builds on the data mining framework by proposing a methodology for analysing sequence events from data in order to determine if there were any unusual occurrences happening on the system and build an investigate profile to help the investigator determine if a particular machine should be considered suspicious.

Meanwhile, Rao et al. (2010) suggests a framework for data analysis (using data mining) that employs a statistical approach to validating the data at the pre-processing stage to ensure the reliability of the data being displayed within the application. The authors state that the proposed model can be adapted to identification of illegally stored data, identification of hidden and encrypted data and identification of renamed file extensions in the future. While digital forensics research into incorporating data mining techniques for large data sets has been limited (Beebe and Clark, 2005), initial studies indicate positive results and highlighting its relevance.

5. Distributed Forensics

The processing power of a single machine does not scale to meet the time-sensitive requirements that are needed to handle most large scale digital forensic operations. Hence, the concept of utilizing the processing power of multiple systems to aid in the handling of a forensics investigation is an attractive proposition for the digital forensics community. Having multiple machines running in parallel and processing the same task (searching or indexing of a large scale dataset, for example) will reduce the overall case investigation time and allow the investigator to focus on what is really important. Indeed, it is easy to see how running tasks such as evidence identification and imaging in parallel would be beneficial in allowing the investigator to focus on acquiring the most relevant data at an early stage of the investigation (Nance et al., 2009).

Van den Hengel et al. (2008) propose a system that uses distributed computing and image processing techniques to quickly extract relevant pieces of data, bringing the important events within the video content to the forefront of the investigator's analysis. The system uses a series of agents to execute tasks on the processing servers, and then return the results to the user in a low-bandwidth format which can be accessed via a pervasive mobile device such as a smartphone.

Roussev et al. (2009) provide a proof of concept that demonstrates how the MPI MapReduce model can be adapted for use in a digital forensic environment for tasks such as indexing, image processing and analytics. The system is based on shared memory that is spread over multiple nodes. The authors argue that by using this model to process forensic related tasks in parallel, terabyte sized datasets can be handled in real-time.

Liebrock et al. (2007) discuss the design of a system that uses parallelism and visual analytics to help reduce the dataset and better handle the imaging of large datasets. Richard and Roussev (2006) discuss a similar distributed processing environment which supports the processing of data in RAM (Random Access Memory) and utilizes more CPU (Central Processing Unit) cycles by spreading the load over multiple nodes. This de-centralized approach helps make investigations more efficient, but does come at a price. Given the need for more hardware, power, associated maintenance and operational aspects, distributed forensics processing incurs higher cost.

AccessData offers a form of parallel data processing in the shape of its PRTK based Distributed Network Attack (DNA) product (AccessData Forensic Toolkit, 2010) by using multiple machines to run the password recovery process in parallel, but more work needs to be done to allow other digital forensics tasks such as imaging, indexing and searching to utilize parallelism over multiple machines and speed up the overall investigation process. Currently AccessData only has multi-core support as part of their Forensic Toolkit (FTK) product (AccessData Forensic Toolkit, 2010) and, while this is a step in the right direction, it is still a way away from a fully-fledged distributed forensic computing environment.

It is worth mentioning that the rapid adoption of cloud computing services as a storage medium is changing the way investigators perceive and deal with digital evidence (Pollitt, 2010). The method and complexity of evidence collection and analysis in a cloud computing environment is a cause for concern. Currently, the digital forensics community lacks the tools required to acquire and analyse data that resides on cloud storage (Pollitt, 2010), which highlights it as an area for future research that is expected to grow in the years ahead.

6. Discussion and Future Work

The following table gives a high-level review of each solution category based on conclusions derived from an extensive literature survey. Future research should involve interviewing practitioners to validate the indications perceived in the literature. An in-depth examination of each implementation is also warranted. Table 1 uses a 3 level rating scale where H = High, M = Medium, and L = Low.

Criterion	Hashing	Triage	Analytics	Distributed Forensics
Accuracy	H	M	M	H
Speed	M	H	M	H
Risk of evidence omission	M	H	M	L
Ease of use	H	M	L	M
Ease of implementation	M	M	M	L

Table 1: Solution Categories Review

Table 1 supports the idea that investigators should not rely on a single technique alone and that the overall solution will probably use a blend of techniques that are dependent on the type of investigation. One area of future research would be to focus on automating forensic analysis tasks over distributed environments. This research should investigate incorporating advanced data analytic techniques and different levels of abstraction to make the investigation of large scale datasets more manageable and less time consuming. A logical order for the implementation of these techniques would be as follows:

1. Identify and assess evidence relevance (triage)
2. Reduce the dataset by removing ‘noisy’ data (hashing and analytics)
3. Process the dataset

Another idea for future research would be to look into the implementation of standard data analytic techniques into triage tools. Is it possible to increase the accuracy of the triage tools with these techniques? These techniques could also be considered for use in data set processing. It would also be interesting to test the Information Retrieval (IR) techniques in a distributed computing environment to see if they are effective. These research areas could be coupled with fuzzy hashing, data fingerprinting or categorization methods to identify and remove similar items from the data set being investigated.

All the proposals put forth thus far are centred on data reduction techniques that attempt to eliminate the 'noisy' data from the dataset so that investigators do not waste valuable time analysing evidence that is of no significance. A system that incorporates multiple such solutions and applies a funnel approach to narrowing the scope of the data would be beneficial. The key to this investigative strategy would be to implement an automated data reduction technique at each level of the funnel in a parallelized fashion, until only the relevant data is revealed. This would allow the investigator to have more thinking time when dealing with the case.

7. Conclusions

As Fred Brooks indicated, there is "no silver bullet" (Brooks, 1987). The same is true for solving the large scale data analysis issue in computer forensics. The increase in time taken and data storage capacities are driving the need for additional research in this area. Industry collaboration with the research community to investigate new and innovative approaches to working with large scale digital storage media for digital forensic investigations would benefit both communities.

This paper discusses the challenges faced by investigators when dealing with large storage media and the different solutions put forth thus far to help combat the problem. These solutions include hashing, fingerprinting, triage, data analytics (searching, categorization, filtering, random data sampling, data mining), and distributed forensics (parallel processing). When using such techniques, there is always going to be a trade-off between missing a possibly crucial piece of evidence versus saving time which, by way of human nature can also be the case even if the investigator manually sifted through each and every bit on the evidence drive.

8. References

Abraham, T. (2006), "Event sequence mining to develop profiles for computer forensic investigation purposes", *In Proceedings of the 2006 Australasian workshops on Grid computing and e-research*, Buyya, R., Ma, T., Rei, S.N., Steketee, C., and Susilo, W. (Eds.), Vol. 54. Australian Computer Society, Inc., Darlinghurst, Australia, 2006, pp. 145-153.

AccessData Website (2011) *AD Triage* (Online). Available from: http://accessdata.com/products/forensic-investigation/ad-triage (Accessed: 29 March 2011)

AccessData Forensic Toolkit (2010) *Sales and Promotional Summary* (Online). Available from: http://accessdata.com/media/en_us/print/techdocs/Forensic%20Toolkit.pdf (Accessed: 29 March 2011)

Apple Website (2011) iPod Products (Online). Available from: http://www.apple.com/uk/ipod (Accessed: 29 March 2011)

Beebe, N. and Clark, J. (2007), "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results", *Digital Investigation*, Vol. 4, Supplement 1, 2007. pp. 49-54.

Beebe, N., Dietrich, G. (2007), "A New Process Model for Text String Searching"*, Advances in Digital Forensics III, IFIP International Federation for Information Processing.* Craiger, P. and Shenoi, S. (Eds.), Vol. 242. National Centre for Forensic Science, Orlando, Florida, 2007. pp. 179-191.

Beebe, N. and Clark, J. (2005), "Dealing with Terabyte Data Sets in Digital Investigations"*, Advances in Digital Forensics, IFIP International Federation for Information Processing*, Pollitt, M. and Shenoi, S. (Eds.), Vol. 194. National Centre for Forensic Science, Orlando, Florida, 2005. pp. 3-16.

Beebe, N. (2009), "Digital Forensic Research: The Good, the Bad and the Unaddressed", *Advances in Digital Forensics V,* Boston, USA, 2009. pp. 17-36.

Brooks, F. P. Jr. (1987), "No silver bullet: essence and accidents of software engineering". Computer, Vol. 20, No. 4, 1987. pp. 0-19.

Carrier, B. (2003), "Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers*", International Journal of Digital Evidence,* IJDE Vol. 1, No. 4, 2003. pp. 1-12.

Carrier, B. (2005), *File System Forensic Analysis*. Addison-Wesley. Indiana, USA. 2005. Chapter 1, p. 6.

Chen, Y., Roussev, V., Richard, G., Gao, Y. (2005), "Content-Based Image Retrieval for Digital Forensics", *Advances in Digital Forensics, IFIP International Federation for Information Processing*, Pollitt, M. and Shenoi, S. (Eds.), Vol. 194. National Centre for Forensic Science, Orlando, Florida, pp. 271-282.

Choudhury, A., Rogers, M., Gillam, B., and Watson, K. (2008), "A Novel Skin Tone Detection Algorithm for Contraband Image Analysis". *In Proceedings of the 2008 Third International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE '08)*. IEEE Computer Society, Washington, DC, USA, 2008, pp. 3-9.

Dabs.com Website (2011) Seagate 3TB FreeAgent GoFlex USB 2.0 3.5" Black Desktop Hard Drive (STAC3000200). Available from: http://www.dabs.com/products/seagate-3tb-freeagent-goflex-usb-2-0-3-5--black-desktop-hard-drive-73QP.html (Accessed April 04 2011)

Dell Website (2011) Optiplex 380 Desktop (Online). Available from: http://www.dell.com/uk/business/p/optiplex-380/pd?oc=x0238004&model_id=optiplex-380&~ck (Accessed: 29 March 2011)

Federal Bureau of Investigation Website (2010) Regional Computer Forensics Laboratory Program Annual Report FY2009 (Online). Available from:

http://www.rcfl.gov/downloads/documents/RCFL_Nat_Annual09.pdf (Accessed 28 March 2011)

Federal Bureau of Investigation Website (2000) Laboratory: Computer Analysis and Response Team (Online). Available from: http://www2.fbi.gov/hq/lab/org/cart.htm (Accessed: 27 March 2011)

Federal Bureau of Investigation Website (2010) Press Release: Muncie Man Sentenced to Seven Years for Distributing Child Porn (Online). Available from: http://indianapolis.fbi.gov/dojpressrel/pressrel10/ip031010.htm (Accessed: 29 March 2011)

Garfinkel, S.L. (2010), "Digital forensics research: The next 10 years", Digital Investigation, Volume 7, Supplement 1, In Proceedings of the Tenth Annual DFRWS Conference, August 2010, pp. 64-73.

Guidance Software EnCase Website (2011) (Online). Available from: http://www.guidancesoftware.com/computer-forensics-fraud-investigation-software.htm (Accessed: 29 March 2011)

IDEAL Technology Corp. Website (2011) STRIKE Overview (Online). Available from: http://www.idealcorp.com/products/index.php?product=STRIKE (Accessed: 29 March 2011)

Lee, S., Bang, J., Lim, K., Kim, J., Lee, S. (2009), "A Stepwise Methodology for Tracing Computer Usage", *NCM '09. Fifth International Joint Conference on Networked Computing, Advanced Information Management and Digital Content and Multimedia Technologies*, 2009. pp. 1852-1857.

Lejsek, H., Þormóðsdóttir, H., Ásmundsson, F., Daðason, K., Jóhannsson, Á., Jónsson, B. and Amsaleg, L. (2010), "Videntifier Forensic: large-scale video identification in practice", *In Proceedings of the 2nd ACM workshop on Multimedia in Forensics, Security and Intelligence (MiFor '10)*, ACM, New York, NY, USA, pp. 1-6.

Liebrock, L.M., Marrero, N., Burton, D.P., Prine, R., Cornelius, E., Shakamuri, M., and Urias, V. (2007). "A preliminary design for digital forensics analysis of terabyte size data sets", *In Proceedings of the 2007 ACM symposium on Applied computing (SAC '07),* ACM, New York, NY, USA, pp. 190-191.

Microsoft Website (2011) Computer Online Forensic Evidence Extractor (COFEE) (Online). Available from: http://www.microsoft.com/industry/government/solutions/cofee/default.aspx (Accessed: 29 March 2011)

Mora, R.J. and Kloet, B. (2010), *Digital forensic sampling* (Online). Available from: http://blogs.sans.org/computer-forensics/files/2010/03/statisticalforensictriage.pdf (Accessed: 27 March 2011)

Nance, K., Hay, B., and Bishop, M. (2009), "Digital Forensics: Defining a Research Agenda", *Proceedings of the 42nd Hawaii International Conference on System Sciences*. 2009. pp. 1-6.

National Software Ref. Library Website (2011) (Online). Available from: http://www.nsrl.nist.gov (Accessed: 30 March 2011)

Oxford Dictionaries Website (2011) 'triage' definition (Online). Available from: http://www.oxforddictionaries.com/definition/triage?view=uk (Accessed: 29 March 2011)

Pollitt, M. (2010), "A History of Digital Forensics", *Sixth Annual IFIP WG 11.9 International Conference on Digital Forensics*. University of Hong Kong, Hong Kong. 2010. pp. 3-15.

Rao, P.G., Bhat, V.H., and Shenoy, D. (2010), "A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application", *IACSIT Int. Journal of Engineering and Technology*, Vol. 2, No. 3 June 2010. pp. 313-319.

Richard, G.G. III and Roussev, V. (2006), "Next-generation digital forensics", *Commun. ACM 49*, 2 (February 2006), pp. 76-80.

Riley, J., Dampier, D., Vaughn, R. (2008). "Time Analysis of Hard Drive Imaging Tools", *Advances in Digital Forensics IV, IFIP International Federation for Information Processing,* Indrajit, R. and Shenoi, S. (Eds.), Vol. 285. National Centre for Forensic Science, Orlando, Florida, 2008, pp. 335-344.

Rogers, M.K., Goldman, J., Mislan, R., Wedge, T., Debrota, S. (2006), "Computer Forensics Field Triage Process Model", *Journal of Digital Forensics, Security and Law*, Vol. 1, No. 2, 2006, pp. 19-38.

Roussev, V. (2009), "Hashing and Data Fingerprinting in Digital Forensics", *Security & Privacy*, IEEE, vol.7, no.2, March-April 2009. pp. 49-55.

Rowlingson, R. (2004), "A Ten Step Process for Forensic Readiness", *International Journal of Digital Evidence*. IJDE Vol. 2, No. 3, 2004. pp. 1-28.

Roussev, V., Wang, L., Richard, G. and Marziale, L. (2009). "A cloud computing platform for large-scale forensic computing", *Advances in Digital Forensics V. Fifth IFIP International Conference on Digital Forensics*, Orlando, Florida, USA, 2009. pp. 201-214.

Spafford, E. (2006), "Some Challenges in Digital Forensics", *Advances in Digital Forensics II*, *IFIP Advances in Information and Communication Technology,* Olivier, M. and Shenoi, S. (Eds.), Vol. 222. National Centre for Forensic Science, Orlando, Florida, 2006, pp. 3-9.

SPEKTOR Forensic Website (2011) Intelligence Forensic Triage (Online). Available from: http://www.evidencetalks.com/spektor.html (Accessed: 29 March 2011)

Sommer, P. (2004), "The challenges of large computer evidence cases", *Digital Investigation*, 1 (1), 2004, pp. 16-17.

Tanner, A.L. and Dampier, D.A. (2010), "An Approach for Managing Knowledge in Digital Forensic Examinations", *International Journal of Computer Science and Security*, (IJCSS), Vol. 4, No. 5, 2010. pp. 451-465.

The Sleuth Kit (2011) (Online). Available from: http://www.sleuthkit.org/sleuthkit (Accessed: 29 March 2011)

US Department of Justice Website (2011) National Drug Intelligence Center: HashKeeper (Online). Available from: http://www.justice.gov/ndic/domex/hashkeeper.htm (Accessed: 29 March 2011)

van den Hengel, A., Hill, R., Detmold, H., and Dick, A. (2008), "Searching in space and time: a system for forensic analysis of large video repositories". *In Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop (e-Forensics '08)*. ICST (Institute for Computer

Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium. pp. 1-6.