ISEEK, A TOOL FOR HIGH SPEED, CONCURRENT, DISTRIBUTED FORENSIC DATA ACQUISITION

Richard Adams[1], Graham Mann[2], Valerie Hobbs[2]
[1]XtremeForensics, [2]Murdoch University
ra@xtremeforensics.com, g.mann@murdoch.edu.au, v.hobbs@murdoch.edu.au

Abstract

Electronic discovery (also written as e-discovery or eDiscovery) and digital forensics are processes in which electronic data is sought, located, secured, and processed with the expectation that it may be used as evidence in legal proceedings. Electronic evidence plays a fundamental role in many aspects of litigation (Stanfield, 2009). However, both eDiscovery and digital forensic approaches that rely on the creation of an index as part of their processing are struggling to cope with the huge increases in hard disk storage capacity. This paper introduces a novel technology that meets the existing and future data volume challenges faced by practitioners in these areas. The technology also addresses the concerns of those responsible for maintaining corporate networks as it does not require installation of 'agents' nor does it have any significant impact on network bandwidth during the search and collection process, even when this involves many computers. The technology is the embodiment of a patented process that opens the way for the development of new functionality, such as the detection of malware, compliance with corporate Information Technology (IT) policies and IT auditing. The technology introduced in this paper has been incorporated into a commercial tool called ISEEK that has already been successfully deployed in a variety of environments.

Keywords: digital forensics; eDiscovery; data acquisition tools; hybrid forensics

INTRODUCTION

Electronic evidence plays a fundamental part in many areas of litigation. In the digital forensic arena, the traditional tools that rely on creating bit-by-bit copy images of devices and then creating an index of their contents are now struggling to cope with the huge increase in hard disk storage capacity seen in recent years (Jusas, Birvinskas, & Gahramanov, 2017). This issue is also present in eDiscovery situations where practitioners typically deal with corporate servers and large numbers of computers (Sondhi, & Arora, 2016). It has therefore become clear that innovation is urgently required if two fundamental aspects of litigation, digital forensics and eDiscovery, are not to impede the legal process through their inability to handle modern volumes of data[i].

Both digital forensics and eDiscovery begin with the forensic acquisition of data that may be used as evidence. For that data to be 'reliable', and therefore admissible, Steel (2006) provides three conditions:

1. The data acquired were from the indicated source
2. The data acquired were collected using proven tools and techniques
3. The data have not been altered since the time of acquisition.

To cope with increasing data volumes, digital forensic practitioners are increasingly resorting to creating 'logical containers' (holding a collection of files and directories) rather than bit-by-bit forensic images. This is very like the collection activity associated with eDiscovery, where the need to process large amounts of data, typically across a network connection, has earned this practice a reputation for being slow, cumbersome and expensive (Sondhi, & Arora, 2014)

Time is always a critical factor in digital forensics and eDiscovery, not only in relation to the court process itself but also in relation to the extent of the disruption caused to those entities involved in collecting data as part of litigation (Adams, Hobbs & Mann, 2013).

In the remainder of this paper, we detail further problems for the collection of electronic data. We discuss current approaches in eDiscovery and digital forensics and identify some of their fundamental limitations. We then propose a new Hybrid Forensics approach to address these problems together with a practical tool called ISEEK. We include some test results from ISEEK deployment in a Windows domain environment and finally summarize

the state of development of ISEEK and comment on other potential uses based on feedback from its deployment in the field.

PROBLEMS WITH TRADITIONAL FORENSIC DATA ACQUISITION

Our experience leads us to believe that, due to the slowness of the process, the creation of bit-by-bit images is practical in only a limited number of cases and that the situation is getting worse given the growth in disk storage capacity (Quick & Choo, 2014), (Franke et al., 2017). The current eDiscovery approach, while also suffering due to the growth in data volumes, is far from being robust. It requires significant human involvement and relies on the creation of indexes that have the potential to miss evidence by, amongst other issues, failing to recognise foreign languages, excluding 'noise words' and by introducing word length restrictions. These issues are covered in more detail under the following headings of Acquisition Speed and Indexing.

Acquisition Speed

In one of our experiments, a series of tests was conducted using virtual machines and virtual disks configured in an 'ideal' situation (i.e. not having to compensate for hardware factors), using a representative selection of forensic imaging tools. These tools were compared for their relative speed to acquire a forensic image of the 160GB source disk[ii]. The results were split into two sections, one section for those tools that could boot into a write-blocked environment (listed in Table 1) and another for those that required some form of separate write-blocking to prevent alteration of the source data (listed in Table 2).

The results shown in Tables 1 and 2 display a wide range of completion times for creation of the forensic images. These results suggest that, assuming the same speeds were maintained, even the fastest tool would take almost two hours to acquire a forensic image of a 1TB disk (Table 1 - IXImager) while the slowest tool would take over 6 hours (Table 2 – EnCase Forensic Imager) even when ignoring issues such as the time to boot a machine or having to remove the disks to attach them to a write-blocking device.

1. Collection Test – Tools Not Requiring Write-Blocker or Dongle

Tool	Time	Image Size	Image Type
IXImager	17 min.	78.6 GB	ASB
Adepto	56 min.	149 GB	RAW
EnCase LineN	1 hr 3 min.	149 GB	EO1
Raptor	1 hr 9min	68.3 GB	EO1

2. Collection Test – Tools Requiring Write-Blocker or Dongle

Tool	Time	Image Size	Image Type
X-Ways Forensic	27 min.	74.4 GB	EO1
FTK Imager	50 min.	149 GB	EO1
EnCase Forensic Imager	1 hr 14 min.	149 GB	EO1

Indexing

In digital forensics, after data acquisition, the next stage is typically to index all the data contained in the forensic image to speed up subsequent searching. When the indexing process was originally developed, the storage capacity of a typical hard disk drive was around 100GB, but now disk drives of 8TB are not uncommon. Unfortunately, the speed of disk storage devices has not kept up with increasing storage capacity meaning that the indexing of a forensic image might take days, even with high-performance processors. In addition, the massive

size increase in the subsequent index files themselves now means that a robust database management system is required to handle them.

Index engines do not recognize (and therefore process) all file types that they come across, and because they tend to determine file type using the file extension they could also be fooled by a malicious user who has changed the file extensions on files they wished to hide.

In addition to having to recognize the file type, because indexes are based around collections of characters (typically words), adding items to an index is only meaningful when it is possible to identify strings of symbols as discrete words or 'related sequences of characters' within a block of source data. That entails not only being able to properly decode all the file types in the data to be indexed, but also managing to identify the words or related sequences of characters contained in those files. This causes problems when faced with foreign languages that do not contain word breaks, i.e. where words do not necessarily have white space characters between them.

To increase speed and reduce index size, indexing algorithms typically ignore white space and 'noise' characters[iii], so the process to retrieve responsive documents may become more complicated in situations where the search term includes either of these features. For example, a sentence such as "Mary had a little lamb" will not exist as a single index entry but is broken up into the separate words requiring the user to find them individually or else put together expressions, for instance by seeking for the word "Mary" within so many words of "lamb". Tools may also place limitations on the length of the words they process to manage the size of the index. Typically, a lower limit of 4 characters is imposed, but this excludes many words that can place a sentence in context. With upper limits set to some arbitrary level, key terms might be excluded such as foreign names, chemical and drug designators. In addition, some words that in English are considered 'noise' words and are therefore excluded from the index may be 'significant' words in another language.

SPECIFIC ISSUES FOR EDISCOVERY

eDiscovery tools typically connect to the 'live' data stored on devices accessible via a network, then create a central index to identify where the data of interest resides. This could be carried out manually with a digital forensics tool, but in an eDiscovery context there are likely to be large numbers of custodians and the requirement for a forensic practitioner to physically set up every instance of data indexing is impractical at the outset, both from the perspective of the amount of time it would take and the logistics of managing a significant number of separate collections.

Many of the leading eDiscovery tools came from existing digital forensics tools that were modified. For instance, Guidance Software added 'agents' to their forensic tool. These agents are small applications that have to be installed across networked systems and that serve as an interface between the central indexing machine and the disks attached to individual computers on a network. These agents also provide a connection to a management system that controls multiple indexing and collection processes. This concept of using agents has been replicated by other developers.

The idea of having agents installed on custodian computers that generate an index and collect data to a central secure point seems logical. However, in practice the fundamental flaws with this approach have become apparent. Notwithstanding all the limitations of an index approach mentioned earlier, the administrators of networks are now reluctant to adopt a process that requires them to install software across these networks, especially when they know that they will cause a large volume of traffic to be generated and could potentially interfere with normal business operations. In addition, given the pressures placed on litigants to meet strict court deadlines, the time required to create an index of data across a large number of systems has become a significant issue for in-house legal teams.

HYBRID FORENSICS APPROACH

The technology exists to overcome the difficulties discussed in previous sections. Rather than imposing restrictions and limits on the search and collection process, it is possible to provide more functionality with greater speed and greatly reduced processing costs.

Traditional approaches to both digital forensics and eDiscovery have focused on a central processing point and have also relied heavily on indexing. With advances in virtualization technology it has been possible to develop an application that runs inside its own virtual environment situated entirely in memory[iv]. This enables the

application to be distributed across an unlimited number of target machines for true parallel processing as each instance is self-contained with no central dependencies. Another advance in technology has provided the ability to search the raw data on a storage device without relying on the operating system to provide access to files, meaning that normally 'locked' files, such as email containers, can be processed. This ability to process email on the custodian machine is a significant benefit given the key role email now plays in litigation.

Combining these two developments provides the ability to carry out parallel processing across a large domain while significantly reducing the volume of data being transported across the network compared to that involved in the 'remote agent and indexing' model.

The Hybrid Forensics approach combines the concept of remote collection of data from multiple sources concurrently (as in eDiscovery) with the collection of the types of data that are generally only important in a digital forensics investigation, e.g. registry information. The key to implementing the Hybrid Forensics approach is an independent collection tool with the ability to undertake literal string searches at a disk level (rather than an operating system level) with the code running entirely in memory on each custodian. This provides five significant benefits:

1. Deployment is fast, easy and doesn't require the participation of custodians
2. Only responsive data is ever moved across the network, thus greatly reducing the impact on the host organisation
3. The search process is much more effective and will find responsive material missed by the index approach
4. The speed of collection is greatly increased as all processing and collection is carried out in parallel rather than individually or in small batches
5. Remote collections on non-networked machines becomes possible.

THE HYBRID FORENSICS APPROACH APPLIED TO DIGITAL FORENSICS

The Hybrid Forensics approach directly addresses three key problems: that of dealing with large data storage devices, acquiring data from multiple systems concurrently and remotely acquiring data with minimum resources at the endpoint.

By design, a bit-by-bit digital forensic image captures the entire contents of a data storage device including deleted and unused space. While it is possible to compress these data, the image is still likely to be too large to be transmitted across a network, especially if more than one image is involved or the data are from a file server or NAS device. The process also requires either that the device is removed from the source computer or that the computer is booted into a digital forensic environment to create the image. Both processes require the hands-on involvement of a digital forensic practitioner.

In some cases, the data of interest can be obtained from a selection of well-defined data types depending on the nature of the investigation, whether these are email (both application-specific and webmail), user files, certain system files (including the registry on Microsoft Windows machines) or deleted files. Hybrid Forensics caters for the collection of all these artefacts. The collected data can be sent to an encrypted container on a device physically attached to the target system while it is still in use. Alternatively, the data can be sent to an encrypted container located on a network share or even to the cloud.

The Hybrid Forensics process can be repeated across as many systems as necessary. These processes run in parallel utilizing the resources of the host systems. Remote collections can be undertaken by:

1. using a deployment agent such as EasyDeploy to run PSExec instances on networked systems that will load and execute the hybrid tool
2. sending a disk containing the hybrid tool plus its configuration file to one or more users at the remote site where it can be replicated and deployed as necessary by a system administrator or consultant with the appropriate access. The data will be sent to a specified target location.
3. sending the hybrid tool plus its configuration file to a system administrator at the remote site who can deploy the tool from a network share and login script or by using PSExec.
4. a system administrator using an RDP session to connect to the remote systems to deploy and run the tool manually.
5. sending a webpage link to selected users for them to download the executable and config files together with instructions for running the tool.

THE HYBRID FORENSICS APPROACH IN PRACTICE

Following the award of a patent for the Hybrid Forensics process[v] an application and its associated configuration and extraction tools have been developed. The suite of tools includes ISEEK-Designer (which creates an encrypted configuration file containing the search/collection containers) and ISEEK-Explorer (which opens the encrypted containers in which are stored the audit results and collected data for viewing and further processing). The deployed search and collection tool itself has been named ISEEK.

The collection process for digital forensic acquisition and that for electronic discovery now appear very similar because with the new methodology the only difference is how the search and collection tool is configured. The key differences between the two applications are:

- For digital forensics, the collected artefacts tend to be complete directories, system files and entire email containers. For eDiscovery, only a limited number of specific files or emails will be collected.
- For digital forensics deleted files are likely to be recovered; these are rarely required for eDiscovery.
- For digital forensics, there will always be data collected from each system, whereas for eDiscovery there may be no items meeting the conditions for collection.

In both digital forensics and eDiscovery collections the configuration files, collected data and logs are encrypted so no aspect of the process is revealed to any unauthorised person who may come into possession of these files.

The ISEEK tool has already been deployed in several instances. In one case, several server farms were searched for data relating to a significant lawsuit in the United States. The entire process was completed with 5 hours whereas a previous attempt using conventional tools was cancelled after several days with no outcome. This case involved searching for terms that were unsuitable for an indexing approach as they included several foreign language terms (including Japanese) coupled with strings of characters that would typically be excluded in an index.

Another case involved a subpoena relating to the emails of 17 bank employees. Following estimates of 3 months to complete the work of identifying relevant emails across approximately 4 TB of data using the existing technology, ISEEK was deployed by two bank employees and within 48 hours they had collected and processed 27,000 relevant emails.

ISEEK is currently being deployed by a large US government contractor, a US military defence agency and a multi-national aerospace company.

TESTING

An experiment to demonstrate the effectiveness of the new process and technology was carried out using 9 custodian systems running a combination of Windows 10 Pro, Windows 10 Enterprise, Windows Server 2012 and Windows Server 2016 in a Windows domain.

Using the deployment utility, nine instances of ISEEK were started on the custodian systems in 48 seconds. ISEEK was configured to locate and collect (to a network share) files and emails containing two search terms that were in the “c:\users” path. The terms were: “Fuld & Company” and “489,628 Dth/d". Both terms are contained in files and attachments to emails within the Enron email data set.

Each custodian system had a mixture of large and small files of various types, including PST, ZIP and HTML. Three of the custodian systems were ‘seeded’ with a PST from the Enron email data set. Each target system had either 2GB RAM (workstations) or 4GB RAM (servers).

The results of the test are shown in Table 3.

3. Results of ISEEK deployment to nine custodian systems in a Windows domain

Machine	Searched data	Responsive data	Number of files searched	Responsive Files	Responsive Emails	Time to complete
WS-1	527 MB	0	3,578	0	0	00:00:45
TRID	14 GB	0	3,772	0	0	00:02:53
WIN-2	9 GB	22 MB	84,259	57	10	00:03:21
WIN-1	13 GB	45 MB	239,017	114	20	00:03:53
ROD	25 GB	0	3,996	0	0	00:05:06
DESK-5	15 GB	0	411,187	0	0	00:08:28
TIG	21 GB	0	6,372	0	0	00:10:18
XF	33 GB	0	9,345	0	0	00:16:38
DESK-4	26 GB	12 MB	548,780	20	4	00:26:12

For comparison purposes, a digital forensics tool that employs an index engine was used to create an index of the same data searched by ISEEK on the custodian system WIN-1. From Table 3 the entire process took just under 4 mins for ISEEK to complete. However, it took 51 mins for the forensics tool to index the same data on the remote system from an i7 8-core system with 12.5GB RAM and creating the index on a local solid-state drive.

In addition, having created an index, the forensics tool was unable to locate the search terms in the same form as that provided to ISEEK, which had completed the whole process on all nine custodian systems in under 30 minutes (with two of those systems containing responsive items processed in under 4 minutes).

Network utilisation peaked at 32Mbps during the process (which included the RDP traffic for monitoring the activities).

Further development is underway to create an integrated deployment application and refine the configuration options by grouping some of them under specific headings, such as the creation of a Forensics tab.

For eDiscovery scenarios, a bulk extraction utility creating XML metadata output together with the collected files for ingesting into a review platform is being refined as well as an API allowing direct access to the encrypted containers for a review platform. A pilot project has already been successful involving the direct import of ISEEK data into the Ringtail review platform.

CONCLUSION

ISEEK has been developed to the stage where it has been used in various environments. The virtualization technology employed has opened the way for the development of further uses with ISEEK, such as processing Windows registry hives for artifacts relevant to security and malware investigations. Conversations with large consulting firms have also identified a potential role for ISEEK in IT compliance engagements, ranging from simply checking licence details of installed software to identifying the presence of confidential documents being stored outside of authorized locations.

Users have identified the key benefits of the Hybrid Technology used in ISEEK as being that:

- the tool does not need to be installed
- the tool does not impact the network infrastructure
- ‘live’ email can be searched and collected without requiring the users to stop working
- the tool can run without the need for any user assistance (or knowledge of the process)
- the process of search and collection is much faster than using alternative methods.

REFERENCES

Adams, R; Hobbs, V. & Mann, G. (2013). The Advanced Data Acquisition Model (ADAM): A process model for digital forensic practice. *Journal of Digital Forensics, Security and Law*, *8*(4): 25-48. doi: https://doi.org/10.15394/jdfsl.2013.1154.

Franke, K. & Årnes, A. (2017). Challenges in digital forensics. In A. Årnes (Ed.), *Digital Forensics* (pp. 313-317). Chichester, England: John Wiley & Sons.

Jusas, V., Birvinskas, D., & Gahramanov, E. (2017). Methods and tools of digital triage in forensic context: Survey and future directions. *Symmetry*, *9*(4), 49. doi:10.3390/sym9040049

Sondhi, S., & Arora, R. (2014). *Applying lessons from e-Discovery to process Big Data using HPC*. Paper presented at the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment, Atlanta, GA, USA. July 13-18, 2014. New York, NY, USA: ACM.

Sondhi, S., & Arora, R. (2016). Big Data processing in the eDiscovery domain. In R. Arora (Ed.), *Conquering big data with high performance computing* (pp. 287-307). Cham: Springer International Publishing.

Stanfield, A. (2009). *Computer forensics, electronic discovery and electronic evidence*. London, England: Reed International Books.

Steel, C (2006). *Windows forensics: The field guide for conducting corporate computer investigations.* Indianapolis, IN: Wiley Publishing.

Quick, D., & Choo, K-K, R. (2014). Impacts of increasing volume of digital forensic data: A survey and future research challenges, *Digital Investigation*, *11*(4) : 273-294. doi: https://doi.org/10.1016/j.diin.2014.09.002

[i] https://www.legaltechnology.com/wp-content/uploads/2013/02/Corporate-Litigation-and-eDisclosure-Current-Trends-and-Future-Challenges.pdf.

[ii] Available at https://www.slideshare.net/RichardAdams3/forensic-imaging-tools-draft-v1-24228558

[iii] https://help.kcura.com/9.2/Content/Recipes/Searching__Filtering__and_Sorting/Using_Stop_Words_and_Making_Some_Characters_Searchable_in_a_dtSearch.htm

[iv] For further information on the technology visit https://turbo.net/

[v] http://www.google.com/patents/US8392706