Digital Investigation 11 (2014) 273–294

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Impacts of increasing volume of digital forensic data: A survey and future research challenges

Darren Quick*, Kim-Kwang Raymond Choo

Information Assurance Research Group, Advanced Computing Research Centre, University of South Australia, Mawson Lakes Campus, Mawson Lakes Boulevard, Mawson Lakes, SA 5095, Australia

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form 1 September 2014
Accepted 4 September 2014
Available online 1 October 2014

Keywords:
Data mining
Data volume
Digital forensics
Evidence discovery
Forensic computer analysis
Intelligence analysis
Knowledge management

ABSTRACT

A major challenge to digital forensic analysis is the ongoing growth in the volume of data seized and presented for analysis. This is a result of the continuing development of storage technology, including increased storage capacity in consumer devices and cloud storage services, and an increase in the number of devices seized per case. Consequently, this has led to increasing backlogs of evidence awaiting analysis, often many months to years, affecting even the largest digital forensic laboratories. Over the preceding years, there has been a variety of research undertaken in relation to the volume challenge. Solutions posed range from data mining, data reduction, increased processing power, distributed processing, artificial intelligence, and other innovative methods. This paper surveys the published research and the proposed solutions. It is concluded that there remains a need for further research with a focus on real world applicability of a method or methods to address the digital forensic data volume challenge.

Introduction

The increase in the number and volume of digital devices seized and lodged with digital forensic laboratories for analysis has been an issue raised over many years. This growth has contributed to lengthy backlogs of work (Gogolin, 2010; Parsonage, 2009). A significant growth in the size of storage media combined with the popularity of digital devices and the decrease in the price of these devices and storage media has led to a major issue affecting the timely process of justice. There is a growing volume of data seized and presented for analysis, often now consisting of many terabytes of data for individual investigations. This has resulted from;

(a) An increase in the number of devices seized per case.

(b) The number of cases with digital evidence is increasing (anecdotal information indicates the last case observed without digital evidence was at least 3 years old).
(c) The size of data on each individual item is increasing.

The increasing number of cases and devices seized is further compounded with the growing size of storage devices (Garfinkel, 2010). Existing forensic software solutions have evolved from the first generation of tools and are now beginning to address scalability issues. However, a gap remains in relation to analysis of large and disparate datasets. Every year the volume of data is increasing faster than the capability of processors and forensic tools can manage (Roussev et al., 2013).

Processing times are increasing with the increase in the amount of data required to be analysed. In the last decade, there have been many calls for research to focus on the timely analysis of large datasets (Garfinkel, 2010; Richard

* Corresponding author. Tel.: +61 8 8172 5074.
E-mail addresses: Darren.Quick@gmail.com, darren.quick@mymail.unisa.edu.au, darren_q@hotmail.com (D. Quick), raymond.choo@unisa.edu.au (K.-K.R. Choo).

http://dx.doi.org/10.1016/j.diin.2014.09.002
1742-2876/

and Roussev, 2006a; Wiles et al., 2007) including the application of data mining techniques to digital forensic data in an endeavour to address the issue of the growing volume of information (Beebe and Clark, 2005; Palmer, 2001).

Serious implications relating to increasing backlogs include; reduced sentences for convicted defendants due to the length of time waiting for results of digital forensic analysis, suspects committing suicide whilst waiting for analysis, and suspects denied access to family and children whilst waiting for analysis (Shaw and Browne, 2013). In addition, employment can be affected for suspects under investigation for lengthy periods of time, and ongoing difficulties can be experienced by suspects and innocent persons when computers and other devices are seized, for example; the child of a suspect may have school assignments saved on a seized computer, or the partner of a suspect may have all their taxation or business information saved on a laptop.

In this paper we study literature examining the digital forensic data volume issue, including the volume of data, the growth of media, and research challenges. We review publications focussing on data mining, data reduction, triage, intelligence analysis, and other proposed methodologies. We then summarise the findings, and future directions for research are outlined in the conclusion.

We located material published in the last 15 years (i.e. 1/1/1999–14/6/2014) by searching various academic databases, including IEEE Xplore, ACM Digital Library, Google Scholar, and ScienceDirect using keywords such as; "Digital Forensic Data Volume", "Computer Forensic Volume Problem", "Forensic Data Mining", "Digital Forensic Triage", "Forensic Data Reduction", "Digital Intelligence", "Digital Forensic Growth", and "Digital Forensic Challenges". In addition, we browsed all papers published in Digital Investigation: The International Journal of Digital Forensics & Incident Response, and The Journal of Digital Forensics, Security and Law. A summary table of key papers and topics is listed in Table 4 (see Discussion section).

Survey

Volume of data (1999–2009)

Digital forensics plays a crucial role in society across justice, security and privacy (Casey, 2014). Concerns regarding the increasing volume of data to be analysed in a digital forensic examinations have been raised for many years. McKemmish (1999) stated that the rapid increase in the size of storage media is probably the single greatest challenge to digital forensic analysis. In 2001, Palmer published the results of the first Digital Forensic Research Workshop (DFRWS), which included a section from Dr Eugene Spafford discussing various challenges posed to computer forensics and stated, 'Digital technology continues to change rapidly. Terabyte disks and decreasing time to market are but two symptoms that cause investigators difficulty in applying currently available analytical tools' (Palmer, 2001).

Sommer (2004) outlined the issues with the increasing data size and number of devices in a legal environment, which is slow to understand the resources and procedures involved, is resulting in methods which do not scale to cope with the increases. Roussev and Richard (2004) stated that the vast amounts of disk storage in use by ordinary computer users would soon overwhelm digital forensic investigators. Ferraro and Russell (2004) discuss the increase ubiquitousness of computers, coupled with a notion of a forensic scientist conducting examinations in every computer related crime, leading to demand for forensic science services which outstrips the resources available, and that alternative methods will be required. Ferraro and Russell (2004) also outline the average time digital evidence is retained, stated to be between three and five years or more, and that orders from courts which can mandate impossible or time consuming procedures in evidence handling, can impede timely processing of evidence. Rogers (2004) reported on a study relating to the needs of digital forensic practitioners, and listed the top issues from a survey conducted of 60 respondents indicating that education, training and certification was the most reported issue, and a lack of funding was the least reported issue, with 'technology' and 'data acquisition' in the top four concerns raised by the respondents.

Brown et al. (2005) stated the challenge in digital forensics is locating relevant information in large datasets, analogous to finding 'needles in haystacks', or in some instances 'bits of needles in bits of haystacks'. Beebe and Clark (2005) state that 'the sheer volume and "noisiness" of … data is absolutely overwhelming and incompatible with manual data analysis techniques.' The unique requirements that make the field of forensic analysis different from traditional pattern analysis include; data that is both related and unrelated, 'interesting' data may be low frequency rather than repetitive, data sources are large and can include multiple sources, differing data types, and that the cost of missing relevant data is large (Brown et al., 2005). Sheldon (2005) stated that due to the increasing capacity of storage devices and the increase in time to undertake analysis, it is 'not feasible to continue performing forensic analysis using the accepted approaches that we use today'.

Alink et al. (2006) state that the volume of data in typical investigations is huge, with modern systems containing hundreds of gigabytes, and large investigations often consist of multiple systems totalling terabytes of data, and in addition, the diversity of data can be overwhelming. Richard and Roussev (2006a) made the observation that most current digital forensic tools are unable to deal with the ever growing size of media, and that new analysis techniques will be required, such as automatic categorisation of pictures. Adelstein (2006) states that the nature of a digital forensic investigation has changed, and the larger disk sizes has resulted in an increase in the time required for collecting a full disk image and then conduct analysis. Furthermore, the nature of digital forensic investigations calls for ongoing technology developments to provide significantly better tools for practitioners (Richard and Roussev, 2006b). As an example, Alink et al. (2006) describe their prototype system, which can display timestamp information merged from different tools, highlighting that current tools, such as EnCase, display time ordered views of file-system metadata only.