Digital Investigation 12 (2015) S22–S29

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2015 Europe

Fast contraband detection in large capacity disk drives

Philip Penrose*, William J. Buchanan, Richard Macfarlane

Edinburgh Napier University, Edinburgh, United Kingdom

Keywords:
Disk sampling
Contraband detection
Digital forensics
Triage
Bloom filter
Sampling
Sample size

ABSTRACT

In recent years the capacity of digital storage devices has been increasing at a rate that has left digital forensic services struggling to cope. There is an acknowledgement that current forensic tools have failed to keep up. The workload is such that a form of 'administrative triage' takes place in many labs where perceived low priority jobs are delayed or dropped without reference to the data itself. In this paper we investigate the feasibility of first responders performing a fast initial scan of a device by sampling on the device itself. A Bloom filter is used to store the block hashes of large collections of contraband data. We show that by sampling disk clusters, we can achieve 99.9% accuracy scanning for contraband data in minutes. Even under the constraints imposed by low specification legacy equipment, it is possible to scan a device for contraband with a known and controllable margin of error in a reasonable time. We conclude that in this type of case it is feasible to boot the device into a forensically sound environment and do a pre-imaging scan to prioritise the device for further detailed investigation.

Introduction

Kryder (2009) shows that the areal density (the number of bits stored per unit area of disk) has been increasing at 40% per year and this is projected to continue for the foreseeable future since the technology is, as yet, nowhere near fundamental limits. This is referred to as 'Kryder's Law' (Walter, 2005) – analogous to Moore's Law for semiconductors. By 2020 it is estimated that a 2.5 inch disk will have a capacity of 14 TB and cost $40. Garfinkel (2010) claims that, because of this, much of the progress made in digital forensic tools over the last decade is becoming irrelevant. These tools were designed to help forensic examiners to find evidence, usually from a relatively low capacity hard disk drive, and do not scale to the capacity of digital storage devices commonly available today. To put this in perspective Roussev et al. (2013) benchmarked the acquisition of a fast 3 TB hard disk drive using a standard acquisition utility at over 11 h. In the UK the Association of Chief Police Officers (ACPO) has acknowledged this situation. Many digital forensics units already have large backlogs and the rate of technological change is likely to accelerate and so exacerbate the situation. Where there is insufficient time or resources to cope with the volume of digital devices being presented a system of forensic triage should be introduced. We will use the term triage to mean a fast initial scan by sampling a digital device, conducted perhaps under severe time and resource constraints, to prioritise the device for possible further detailed investigation. Young et al. (2012) argue that it is critical for forensic investigators to have such a triage process so that they can quickly detect bad or illegal files on a large disk. While consulting with digital forensic analysts from Police Scotland it was queried whether it was feasible to implement such a forensic triage tool directly on a suspect machine so that a responding officer could quickly assess whether it held any contraband. The requirements would be that the system should:

* Corresponding author.
E-mail address: p.penrose@napier.ac.uk (P. Penrose).

http://dx.doi.org/10.1016/j.diin.2015.01.007

- Be 99.9% accurate
- Give results in a reasonable time
- Execute on low specification legacy equipment.

Restrictions on our methodology are imposed by these requirements. To achieve results in a reasonable time we must sample the device rather than inspect every sector. If we sample complete files on a disk then the file metadata is read from the file system and then the file is accessed. This involves considerable physical head movement in the disk drive. Fujitsu (Fujitsu Technology Solutions GmbH, 2011) benchmarked a fast hard disk drive which showed random access throughput at 3 MB/s and sequential access at 200 MB/s. Thus, random sampling of files would be considerably slower for triage purposes. In addition, it relies on the file system metadata and so would not cover unallocated space on the disk where illicit material may well be hidden. Statistically sampling disk sectors overcomes both problems. If the random selection of sector addresses is sorted before accessing the disk, then this incurs only a single sequential pass over the disk since the disk is treated as simply a sequence of blocks. Additionally, since the sample is chosen from the whole address range of the disk it would sample all areas including unallocated space. Also, using sector sampling bypasses the file system by sampling raw disk sectors and so the nature of the file system on the disk is irrelevant.

Another consequence of the restrictions is that any reference data set should be held in RAM so that lookups can be done to match disk read speed. Lookup of a disk based database would be slower than the sequential sampling. Hence a compact representation of any reference data set is needed if we are to legacy equipment with limited RAM.

Previous work

Triage

Pollitt (2013) argues that the process of digital forensic triage is an admission of failure. The backlog of cases is due to the systemic failure of the digital forensic process and of digital forensic software. These have not adapted to the vast increase in digital data that is involved in a modern case. Triage has become necessary because investigators often prefer some useful evidence quickly rather than wait, perhaps some considerable time, for all detectible evidence to be found. He argues that by focussing on a particular outcome such as the existence of specific types of data, we miss important information, such as logs or e-mail that might reveal a wider group of suspects. However, we argue that triage must not be the only tool used in an investigation. If any incriminating evidence is found during the process of triage then the device should be subject to a full digital forensic analysis.

Shaw and Browne (2013) note that '*administrative*' triage already takes place in many organisations and criteria are used to either prioritise or exclude a device from examination. Horsman et al. (2014) maintain that organisations may also be cautious because there is a perception that there is a risk of missing evidence where triage only samples a device. We argue that in our scenario of detecting contraband, there is less risk in a system of triage allowing the timely analysis of a device using forensically sound boot media, and with a controllable probability of missed evidence, than a system of administrative triage which operates without any reference to the physical media.

Existing triage solutions

There are a number of triage packages available, both open source and proprietary, such as Strike, EnCase Portable, AD Triage, Triage IR, Kludge. These packages typically perform data collection, often with no intervention by the operator, of such things as internet history, the registry, file metadata, recently used files, image files, hash lookup of user home directory files and comparing to a selected file hash database, indexing, keyword matching and so on. Some even do full disk imaging as part of triage. File carving from unallocated disk space is also an option on some. They uniformly behave as versions of full forensic analysis packages, collecting appropriate data for later examination. Most are designed so that they can be used by an untrained operative and give on-screen display of images or analysis results as they are produced. Casey et al. (2013) view this type of product as freeing forensic analysts from the routine task of acquiring forensic evidence and empowering them to concentrate on the more interesting aspects of their work. However these tools should be regarded as automating the acquisition stage of a forensic investigation rather than as triage. All the operations performed tend to be I/O and processor intensive and defeat the purpose of our definition of triage – a fast initial scan to ascertain if a device contains images or documents of interest. Gillam and Rogers (2005) developed File Hound, a forensic tool for first responders. This application was file based and would not be suitable for direct sector sampling. It reported all images found and did not use a reference data set. Roussev et al. (2013) treat triage as an intrinsic part of the digital forensic process. They advocate that target acquisition and forensic processing should be done in parallel, with results being reported as soon as they are available. Their model requires that data is analysed as it is being acquired so that analysis should finish at the same time as the data acquisition. This requires that analysis, including cryptographic hashing and lookup, similarity hashing, decompression, file content extraction and indexing all be done in parallel at the speed of data acquisition. To do this needs considerably more computing power than is available in the field. We would argue that there is still a demonstrated need for a separate initial triage stage prior to the computing power for this acquisition and forensic processing becoming available to the investigator. Garfinkel (2013) developed *bulk_extractor* to scan an entire disk image. A disk image is scanned without reference to partitions or the file system metadata. Since this method does not have to find, extract identify and process *files*, it is shown to be at least ten times faster than traditional file based methods. It used a number of filters running in parallel, each optimised to detect patterns