Contents lists available at ScienceDirect

Digital Investigation

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

Monte-Carlo Filesystem Search – A crawl strategy for digital forensics

Janis Dalins [a], [b], [*], Campbell Wilson [a], Mark Carman [a]

[a] Faculty of Information Technology, Monash University, 900 Princes Highway, Caulfield East, Victoria, Australia
[b] Digital Forensics, Australian Federal Police, 383 La Trobe Street, Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 4 February 2015
Received in revised form 31 March 2015
Accepted 1 April 2015
Available online 25 April 2015

Keywords:
Monte-Carlo Filesystem Search
Search warrant
Triage
Computer forensics
PhotoDNA
Skin tone
Crawl strategy
Monte-Carlo
Child pornography

ABSTRACT

Criminal investigations invariably involve the triage or cursory examination of relevant electronic media for evidentiary value. Legislative restrictions and operational considerations can result in investigators having minimal time and resources to establish such relevance, particularly in situations where a person is in custody and awaiting interview. Traditional uninformed search methods can be slow, and informed search techniques are very sensitive to the search heuristic's quality. This research introduces Monte-Carlo Filesystem Search, an efficient crawl strategy designed to assist investigators by identifying known materials of interest in minimum time, particularly in bandwidth constrained environments. This is achieved by leveraging random selection with non-binary scoring to ensure robustness. The algorithm is then expanded with the integration of domain knowledge. A rigorous and extensive training and testing regime conducted using electronic media seized during investigations into online child exploitation proves the efficacy of this approach.

Introduction

Digital forensic analysis relies upon the investigator(s) being able to access relevant data in a lawful and timely fashion. In criminal investigations across most common law jurisdictions, lawful seizure of electronic media and/or data is achieved through the provisions of a search warrant issuedin accordance with applicable law. Search warrants in Australian Commonwealth criminal investigations are often issued under Section 3E of the *Crimes Act 1914* (Cth), giving investigators permission to search premises, conveyances, and/or person(s) and seize evidential materials (subject to meeting specific criteria). To paraphrase the relevant legislation, electronic items can only be seized if the investigator believes on reasonable grounds that the item, or data accessed by operating the item, is evidential material.[1]

An investigator executing a search warrant is therefore faced with the challenge of examining all potentially relevant electronic devices within a target premises before being able to seize items and/or copy data. Typically, this analysis will involve manual browsing of data (perhaps with the targeting of specific features), or the calculation of file hashes with subsequent comparisons against pre-established hash sets of known files. Both processes are resource intensive from a computational,

[*] Corresponding author. Faculty of Information Technology, Monash University, 900 Princes Highway, Caulfield East, Victoria, Australia.

E-mail addresses: janis.dalins@afp.gov.au, janis.dalins@monash.edu (J. Dalins), campbell.wilson@monash.edu (C. Wilson), mark.carman@monash.edu (M. Carman).

[1] Section 3K *Crimes Act 1914* (Cth) allows for the temporary moving of items offsite for examination, but this is subject to time restrictions and other considerations.

http://dx.doi.org/10.1016/j.diin.2015.04.002
1742-2876/© 2015 Elsevier Ltd. All rights reserved.

bandwidth, and/or human perspective. The impact of inefficient search is increased due to both practical and legislative reasons:

- **Practical** – Examinations are carried out on premises, often using suspects' hardware. The safety and security of such situations can vary widely and unpredictably, as can the quality and performance of available infrastructure; and
- **Legislative** – Australian Commonwealth legislation[2] places strict limits (typically 4 h) on the time allowed between arrest and laying of charges or release, making any increase in relevant information available to investigators during this time extremely valuable.

This research proposes moving away from existing methods of arbitrary file system structure 'walks'. Instead, we propose the Monte-Carlo Filesystem Search (MCFS) algorithm for efficient file system search through rigorous prioritisation of files and directories for examination. In this paper, we illustrate the method in the context of discovering files known to contain child pornography. We continue the established method of using MD5 hash values as ultimate file identifiers, but also provide our crawler with richer feedback through similarity scoring (PhotoDNA) and skin tone detection within otherwise unknown image files. The modular nature of MCFS makes it suitable for use with any ranking algorithm.

We emphasise that this research and the methods proposed within this paper are not means for *complete* examination of electronic media for all potential items of interest. They are intended to serve as a means for supporting triage by expediting the identification of known files of interest, guiding subsequent in-depth investigation.

The contribution of this paper is the design, introduction and evaluation of MCFS, a digital forensic crawl strategy predicated upon:

- **Speed:** Searches must be as fast as practicable, with scores readily (if not immediately) available throughout the runtime;
- **Efficiency:** The method(s) must be lightweight from a memory and bandwidth perspective; and
- **Accuracy:** The method(s) must not introduce unacceptably high levels of false positive (and particularly false negative) results.

We found that whilst MCFS is effective using a simple binary identification method (MD5 hashing), it excels when provided with finer scoring granularity. Performance improvements were observed using skin tone detection, with image similarity (using PhotoDNA) found to greatly improve search efficiency.

This paper is structured as follows. First we provide a brief overview of MCFS, and discuss the desirable characteristics of a forensic search tool based on the algorithm. We then examine the use of domain specific knowledge regarding the filesystem search context to customise the algorithm for forensic analysis. This is followed by a discussion of the results of experiments designed to demonstrate the efficacy of this approach under various forensic scenarios.

Related work

The impact of rapidly growing data quantities within digital forensics has long been identified. Numerous methods for overcoming the corresponding increases in time and resource requirements have been proposed. Our approach overlaps with several such categories of methods, as defined by Quick and Choo (2014): *triage*, *data mining*, and to a lesser extent, *data reduction and subsets*.

Triage

Broadly speaking, digital forensic triage (defined by Roussev et al. (2013) as *"a partial forensic examination conducted under (significant) time and resource constraints"*) is the closest relation to MCFS, due to our focus upon rapid identification of data for investigator use within interviews and early investigations.

Roussev and Quates (2012) identify the slow performance of 'deep forensic' examinations, instead choosing to focus upon the use of similarity digests as a means for identifying correlations across sources and establishing "an initial framework of understanding" for studied cases. The authors identify metadata based prioritisation as an option for improving performance, but instead gain a performance advantage through sequential access to the physical storage device.

An interesting analysis of performance impacts is presented by Roussev et al. (2013), who perform typical investigative tasks on a reference target using 'workstation' and 'server' configurations, reflecting on-site and lab-based triage. Whereas metadata extraction and cryptographic hashing (e.g. MD5) perform well on the workstation, more intensive methods such as indexing and similarity hashing 'become somewhat feasible on the server'.

Data mining

Machine learning has been proposed for use in triage, with Marturana and Tacconi (2013) building classifiers based upon investigation type specific matrices of features (for example, number of installed P2P applications in a copyright infringement matter) on suspect devices. The classifiers are then used to provide analysts with a rapid indication of the device's likely relevance to the matter at hand.

da Cruz Nassif and Hruschka (Jan 2013) research the efficacy of clustering within digital forensics, using partitional (K-means, K-medoids), hierarchical (Single Link, Complete Link, Average Link) and cluster ensemble (CSPA) algorithms on real-world datasets from Brazilian Federal Police investigations. Average Link and Complete Link algorithms performed best on the datasets, with "suitably initialized" K-means and K-medoids also performing well.

[2] *Crimes Act 1914* (Cth) Section 23C.