available at www.sciencedirect.com

ScienceDirect

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

Digital Investigation

Digital forensics research: The next 10 years

Simson L. Garfinkel

Naval Postgraduate School, Monterey, USA

Keywords:
Forensics
Human subjects research
Corpora
Real data corpus
Realistic data

ABSTRACT

Today's Golden Age of computer forensics is quickly coming to an end. Without a clear strategy for enabling research efforts that build upon one another, forensic research will fall behind the market, tools will become increasingly obsolete, and law enforcement, military and other users of computer forensics products will be unable to rely on the results of forensic analysis. This article summarizes current forensic research directions and argues that to move forward the community needs to adopt standardized, modular approaches for data representation and forensic processing.

1. Introduction

Digital Forensics (DF) has grown from a relatively obscure tradecraft to an important part of many investigations. DF tools are now used on a daily basis by examiners and analysts within local, state and Federal law enforcement; within the military and other US government organizations; and within the private "e-Discovery" industry. Developments in forensic research, tools, and process over the past decade have been very successful and many in leadership positions now rely on these tools on a regular basis—frequently without realizing it. Moreover, there seems to be a widespread belief, buttressed on by portrayals in the popular media, that advanced tools and skillful practitioners can extract actionable information from practically any device that a government, private agency, or even a skillful individual might encounter.

This paper argues that we have been in a "Golden Age of Digital Forensics," and that the Golden Age is quickly coming to an end. Increasingly organizations encounter data that cannot be analyzed with today's tools because of format incompatibilities, encryption, or simply a lack of training. Even data that can be analyzed can wait weeks or months before review because of data management issues. Without a clear research agenda aimed at dramatically improving the efficiency of both our tools and our very research process, our hard-won capabilities will be degraded and eventually lost in the coming years.

This paper proposes a plan for achieving that dramatic improvement in research and operational efficiency through the adoption of systematic approaches for representing forensic data and performing forensic computation. It draws on more than 15 years personal experience in computer forensics, an extensive review of the DF research literature, and dozens of discussions with practitioners in government, industry, and the international forensics community.

1.1. *Prior and related work*

Although there has been some work in the DF community to create common file formats, schemas and ontologies, there has been little actual standardization. DFRWS started the Common Digital Evidence Storage Format (CDESF) Working Group in 2006. The group created a survey of disk image storage formats in September 2006, but then disbanded in August 2007 "because DFRWS did not have the resources required to achieve the goals of the group. (CDESF working group, 2009)" Hoss and Carver discuss ontologies to support digital forensics (Carver and Hoss, 2009), but did not propose any concrete ontologies that can be used. Garfinkel introduced an XML representation for file system metadata (Garfinkel, 2009), but it has not been widely adopted.

E-mail address: simsong@acm.org

1742-2876/$ – see front matter © 2010 Digital Forensic Research Workshop. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2010.05.009

Richard and Roussev reviewed requirements for "Next-generation digital forensics." Their work stressed system requirements, and argued that inefficient system design, wasted CPU cycles, and the failure to deploy distributing computing techniques is introducing significant and unnecessary delays that directly translate into unnecessary delays (Richard and Roussev, 2006). Elements of a modular computer forensics system exist in both Corey et al.'s design of a network forensics analysis tool (Corey et al., 2002) and in Cohen's PyFlag (Cohen, 2008), although the rest of the DF research community has generally failed to appreciate how these architectures can satisfy Richard and Roussev's requirement for parallelism. Ayers ignored all of the previous work on this topic in his "second generation computer forensic analysis system," presented at DFRWS 2009 (Ayers, 2005). In general, it seems that very few DF systems designers build upon previous work—instead, each new project starts afresh.

Following the first DFRWS, Mocas proposed a framework to help build "theoretical underpinnings for digital forensics research (Mocas, 2004)." The purpose of the framework was to "define a set of properties and terms that can be used as organizing principles for the development and evaluation of research in digital forensics." Mocas suggested that research should consider context in which evidence is encountered, data integrity, authentication, reproducibility, non-interference and the ability of proposed techniques to comply with federal minimization requirements.

Pollitt reviewed 14 different models for digital forensics investigation but did not attempt to evaluate or catalog them given time constraints (Pollitt, 2007). Most of these investigation models rely on the ability to make the best use of digital evidence that is found. An alternative approach is *proactive digital forensics*—for example, Ray et al.'s design for a system that predicts attacks and changes its collection behavior *before* an attack takes place (Allen Ray, 2007). Bradford et al. likewise argue that it is unwise to depend upon "audit trails and internal logs" and the digital forensics will only be possible on future systems if those systems make proactive efforts at data collection and preservation; they present a mathematical model for deciding the content and frequency of proactive forensic event recorders (Bradford et al., 2004).

Pollitt et al. discussed how virtualization software and techniques can be productively applied to both digital forensics research and education (Pollitt et al., 2008). Any discussion of virtualization with respect to digital forensics faces an unwelcome tautology. In practice, the impact of virtualization on forensic examination can usually be ignored—except when it can't. That's because sometimes the virtualization is the subject of the forensic examination, and sometimes the virtualization is a tool it is used by the forensic examiner.

In June 2008 a brainstorming session at CISSE 2008 explored research categories, topics and problems in digital forensics. One of the results of this project was an article by Nance, Hay and Bishop that attempted to define a Digital Forensics Research Agenda (Nance et al., 2009). The authors identified six categories for digital forensics research: Evidence Modeling, Network Forensics, Data Volume, Live Acquisition, Media Types, and Control Systems. This taxonomy is useful, but believe that the tactical analysis must be accompanied by strategic thinking.

In January 2009 Beebe presented an invited talk at the Fifth IFIP WG 11.9 International Conference on Digital Forensics entitled "Digital Forensics: The Good, The Bad, and the Unaddressed (Beebe, 2009)." Beebe argued convincingly that digital forensics was no longer a niche discipline. "It is now mainstream knowledge that the digital footprints that remain after interactions with computers and networks are significant and probative. Digital forensics was once a niche science that was leveraged primarily in support of criminal investigations, and digital forensic services were utilized only during the late stages of investigations after much of the digital evidence was already spoiled. Now digital forensic services are sought right at the beginning of all types of investigations…Even popular crime shows and novels regularly incorporate digital evidence in their story lines."

As far as "The Bad" and "The Unaddressed," Beebe said that digital forensics largely lacks standardization and process, and what little widespread knowledge that we have is "heavily biased towards Windows, and to a lesser extent, standard Linux distributions." Unaddressed, Beebe says, is the problem of scalability, the lack of intelligent analytics beyond full-text search, non-standard computing devices (especially small devices), ease-of-use, and a laundry list of unmet technical challenges.

Finally, Turnbull et al. performed a detailed analysis on the specific digital media formats being collected by the South Australian Police Electronic Crime Section; theirs appears to be the first quantitative analysis of its kind (Turnbull et al., 2009), although the FBI's Regional Computer Forensic Laboratory program publishes an annual report with the amount of media and cases that it processes (Regional Computer Forensics Laboratory, 2008). More case studies such as these are needed so that researchers can use actual evidence, rather than their own personal experiences, to direct their problem-solving efforts.

2. Digital forensics: a brief history

Today DF is an important tool for solving crimes committed with computers (*e.g.* phishing and bank fraud), as well as for solving crimes against people where evidence may reside on a computer (*e.g.* money laundering and child exploitation). Forensic tools have also become a vital tool for Information Assurance because of their ability to reconstruct the evidence left by cyber attacks.

2.1. *The early days*

DF is roughly forty years old. What we now consider forensic techniques were developed primarily for data recovery. For example, Wood et al. relate a story about two local data recovery experts working for 70 h to recover the only copy of a highly fragmented database file inadvertently erased by a careless researcher (pp.123–124 Wood et al., 1987). By the late 1980s utilities were being widely advertised that could perform a variety of data recovering, including "Unformat, Undelete, Diagnose & Remedy"(p.57 Display ad 57, 1987).

These early days were marked by: