Digital Investigation 13 (2015) 94–106

Contents lists available at ScienceDirect

Digital Investigation

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

Automated event and social network extraction from digital evidence sources with ontological mapping

Benjamin Turnbull [a, *], Suneel Randhawa [b]

[a] Australian Centre for Cyber Security, University of New South Wales, ACT, Australia
[b] Automated Analytics and Decision Support, Cyber and Electronic Warfare Division, Defence Science and Technology Organisation, West Avenue, Edinburgh, South Australia, 5111, Australia

ARTICLE INFO

Article history:
Received 11 July 2013
Received in revised form 22 April 2015
Accepted 26 April 2015
Available online 18 May 2015

Keywords:
Artificial intelligence
Big data
Digital forensics
Digital evidence
Event representation
Forensic tool development
Knowledge representation
Ontology
Software engineering
Triage

ABSTRACT

The sharp rise in consumer computing, electronic and mobile devices and data volumes has resulted in increased workloads for digital forensic investigators and analysts. The number of crimes involving electronic devices is increasing, as is the amount of data for each job. This is becoming unscaleable and alternate methods to reduce the time trained analysts spend on each job are necessary.

This work leverages standardised knowledge representations techniques and automated rule-based systems to encapsulate expert knowledge for forensic data. The implementation of this research can provide high-level analysis based on low-level digital artefacts in a way that allows an understanding of what decisions support the facts. Analysts can quickly make determinations as to which artefacts warrant further investigation and create high level case data without manually creating it from the low-level artefacts. Extraction and understanding of users and social networks and translating the state of file systems to sequences of events are the first uses for this work.

A major goal of this work is to automatically derive 'events' from the base forensic artefacts. Events may be system events, representing logins, start-ups, shutdowns, or user events, such as web browsing, sending email. The same information fusion and homogenisation techniques are used to reconstruct social networks. There can be numerous social network data sources on a single computer; internet cache can locate Facebook, LinkedIn, Google Plus caches; email has address books and copies of emails sent and received; instant messenger has friend lists and call histories. Fusing these into a single graph allows a more complete, less fractured view for an investigator.

Both event creation and social network creation are expected to assist investigator-led triage and other fast forensic analysis situations.

Introduction

If you ask any digital forensic analyst or manager over the last decade, one of the greatest organisational issues faced in electronic evidence analysis is the growth of workload. The last decade of Moore's and Kryder's Laws (Schaller, 1997; Walter, 2005) and the proliferation of devices in an approaching 'post-pc' world has seen computing encroach upon every facet of our lives. The growth in what defines a technology as a suitable source for forensic analysis and the amount of data each component may hold has grown substantially and electronic crime and forensic analysts are straining under the weight of this demand and

* Corresponding author.
E-mail addresses: Benjamin.Turnbull@unsw.edu.au (B. Turnbull), Suneel.Randhawa2@dsto.defence.gov.au (S. Randhawa).

http://dx.doi.org/10.1016/j.diin.2015.04.004
1742-2876/

are trying to find new ways of coping with the increasing influx of computer data for analysis.

Given the growth of workload is not changing in the near future, there is a need to augment the methods being employed by forensic analysis groups to increase the number of devices that can be analysed without sacrificing the quality of results. We need to work smarter, not harder; or where possible, employ automated services to perform some of this work.

There are two primary aims of this research; to provide a large-scale, consistent knowledge representation and to build symbolic Artificial Intelligence for developing deep understanding in digital forensic cases based on that data. Specifically, by encoding the data into ontology and reasoning over the resulting dataset, higher abstractions of data can be derived. Although other AI and machine learning paradigms are of use in this field, our current focus is on systems that provide the ability to audit the decision making process of inferred knowledge. Extraction of users and social networks and translating the state of file systems to sequences of events are the first two uses for this work.

The first outcome of this work is an ontological representation and data store consistently representing entities and relationships pertaining to:

- The hierarchy of files, directories and file systems,
- User accounts and system information,
- System events,
- People,
- User events.

The second outcome of this work is a rule-based system that automatically extracts data from multiple sources. One of the major goals of this work is to automatically infer 'events' from the base forensic artefacts. Events may be system events (taken from multiple sources and fused appropriately), representing logins, start-ups, shutdowns and updates, or user events, such as web browsing, sending email or other activity. The same information fusion and homogenisation techniques can also be used to reconstruct social networks. There are numerous social network data sources located on a single computer; internet cache can house Facebook, LinkedIn, Google Plus caches; email has address books and copies of emails sent and received; instant messenger has friend lists and call histories. Fusing these into a single graph allows a more complete, less fractured view for an investigator, providing better insight.

This work is then implemented as a proof-of-concept forensic tool, *ParFor* ('Parallax Forensics'). *ParFor* makes use of Resource Descriptive Framework (RDF) (Klyne et al., 2004) ontologies to provide a unified representation of multiple different data sources, and to provide higher level reasoning capabilities. *ParFor* was designed primarily as a vehicle for this research, but was scoped in a relatively generic manner. As such, it is expected that this platform is generic enough to serve as a basis for other machine learning and reasoning paradigms, as well as allow expansion into other forms of digital evidence. This work is conducted as part of the Parallax BattleMind project (Murray et al., 2013).

Background and research need

Given the increased workload of digital forensic analysts everywhere, investigator led triage has become a common method for moving work from overloaded forensic analysts to investigators. It is thought that an investigator with some computer training, rather than a specialist forensic analyst, has more understanding of the wider investigation and is often best-suited to perform an initial analysis on a device. This is especially true for cases where the main purpose of the analysis is to extract specific information, or where an investigator is primarily interested in a 'summary' of the device to see if further analysis is necessary by analysts. Expert forensic analysts can therefore be diverted to more technical analysis, making more efficient use of their time and expertise. There are disadvantages to the use of triage as a substitute for analysis (Pollitt, 2013), but that is beyond the scope of this work. The reality is that triage is a commonly used technique and this is unlikely to change in the near future.

Assistance to less technical investigators can take several forms, such as additional training. This approach has had several success stories (Schmidt et al., 2009; Dampier et al., 2012). One alternative is to alter the software to make it friendlier for less technical investigators or to construct specific tools that can encode expert knowledge. Currently, most forensic systems and tools, even triage systems, are designed for technical forensic analysts. While triage-specific tools exist, often in practice investigators use full forensic analysis software for triage but only making use of a reduced feature set. Reducing the need to locate and interpret low-level computing artefacts is another, less explored method for achieving the same goals.

The core concept of this approach is that expert systems can locate and interpret the artefacts and provide higher level abstractions/concepts vs. conclusions. As long as the reasoning behind the reasoning process can be explained, either for court or for verification purposes, such an approach can be used to interpret a piece of digital evidence whilst abstracting away the underlying mechanics of file interpretation.

One logical abstraction of computing devices is to define it as a sequence of events that have occurred on the system. For example, an investigator wants to know when a user was accessing the internet and what sites they visited; they care less about the browser and operating system specific encoding of how this information is stored. It is all about the interpretation, not the files themselves. It is less about what is there and more about what it means.

In this sense, this work expands upon the work of Carrier (Carrier, 2003), which classifies individual tools at multiple abstraction levels, providing the abstraction layers *physical media*, *media management*, *file system analysis*, *network analysis* and *memory analysis*. This work increases the abstraction layers beyond those of Carrier and into the user space. The concepts, however, remain the same. Each level is unique and some levels of abstraction may build upon the interpretation of the previous ones. As long as the link between abstraction layers is understood, the higher