Digital Investigation 11 (2014) S95–S105

Contents lists available at ScienceDirect

Digital Investigation

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

A complete formalized knowledge representation model for advanced digital forensics timeline analysis

Yoan Chabot [a, b, *], Aurélie Bertaux [a], Christophe Nicolle [a], M-Tahar Kechadi [b]

[a] CheckSem Team, Laboratoire Le2i, UMR CNRS 6306, Faculté des sciences Mirande, Université de Bourgogne, BP47870, 21078 Dijon, France

[b] School of Computer Science & Informatics, University College Dublin, Belfield, Dublin 4, Ireland

Keywords:
Digital forensics
Timeline analysis
Event reconstruction
Knowledge management
Ontology

A B S T R A C T

Having a clear view of events that occurred over time is a difficult objective to achieve in digital investigations (DI). Event reconstruction, which allows investigators to understand the timeline of a crime, is one of the most important step of a DI process. This complex task requires exploration of a large amount of events due to the pervasiveness of new technologies nowadays. Any evidence produced at the end of the investigative process must also meet the requirements of the courts, such as reproducibility, verifiability, validation, etc. For this purpose, we propose a new methodology, supported by theoretical concepts, that can assist investigators through the whole process including the construction and the interpretation of the events describing the case. The proposed approach is based on a model which integrates knowledge of experts from the fields of digital forensics and software development to allow a semantically rich representation of events related to the incident. The main purpose of this model is to allow the analysis of these events in an automatic and efficient way. This paper describes the approach and then focuses on the main conceptual and formal aspects: a formal incident modelization and operators for timeline reconstruction and analysis.

Introduction

Due to the rapid evolution of digital technologies and their pervasiveness in everyday life, the digital forensics field is facing challenges that were anecdotal a few years ago. Existing digital forensics toolkits, such as EnCase or FTK, simplify and facilitate the work during an investigation. However, the scope of these tools is limited to collection and examination of evidence (i.e. studying its properties), which are the first two steps of the investigation process, as defined in Palmer (2001). To extract acceptable evidence, it is also necessary to deduce new knowledge such as the causes of the current state of the evidence (Carrier and Spafford, 2004b). The field of event reconstruction aims at solving this issue: event reconstruction can be seen as a process of taking as input a set of events and outputting a timeline of the events describing the case. Several approaches have been proposed to carry out event reconstruction, which try to extract events and then represent them in a single timeline (super-timeline (Gudhjonsson, 2010)). This timeline allows to have a global overview of the events occurring before, during and after a given incident. However, due to the number of events which can be very large, the produced timeline may be quite complicated to analyse. This makes the interpretation of the timeline and therefore the decision making very difficult. In addition, event reconstruction is a complex

* Corresponding author. CheckSem Team, Laboratoire Le2i, UMR CNRS 6306, Faculté des sciences Mirande, Université de Bourgogne, BP47870, 21078 Dijon, France.

E-mail address: yoan.chabot@checksem.fr (Y. Chabot).

http://dx.doi.org/10.1016/j.diin.2014.05.009
1742-2876/

process where each conclusion must be supported by evidence rigorously collected, giving it full credibility.

In this paper, we first address these problems by proposing an approach to reconstruct scenarios from suspect data and analyse them using semantic tools and knowledge from experts. Secondly, this paper answers the challenge of correctness of the whole investigative process with a formal incident modelization and timeline reconstruction and analysis operators. The paper is organized as follows. Section Related Works reviews important issues of the events reconstruction problem and the various approaches proposed so far. The SADFC (*Semantic Analysis of Digital Forensic Cases*) approach is described in Section SADFC approach, and the formal advanced timeline reconstruction and analysis model is presented in Section Advanced timeline analysis model. Finally, a case study illustrating the key characteristics of the proposed approach is given in Section Case study.

Related works

Events reconstruction has many issues, which are directly related to the size of the data, digital investigation process complexity, and IT infrastructures challenges. For instance, Table 1 compares some existing approaches (their strengths (✓), limitations (✖), partial or inadequate solutions (●)) with regard to some key issues, such as heterogeneity, automatic knowledge extraction, the use of proven theory as support, analysis capabilities, and preservation of data integrity. While some of these challenges have been a focus for many researchers and developers for the last decade, the size of data volumes (Richard III & Roussev, 2006) and data heterogeneity are still very challenging. The first (large data sizes) introduces many challenges at every phase of the investigation process; from the data collection to the interpretation of the results. The second (data heterogeneity) is usually due to multiple footprint sources such as log files, information contained in file systems, etc. We can classify events heterogeneity into three categories:

- *Format*: The information encoding is not the same among sources due to the formatting. Therefore, depending on the source, footprint data may be different.
- *Temporal*: The use of different sources from different machines may cause timing problems. First, there are some issues due to the use of different time zones and unsynchronized clocks. Second, the temporal heterogeneity can be due to the use of different formats or granularities (e.g. 2 s in FAT file systems, 100 ns in NTFS file systems).
- *Semantic*: The same event can be interpreted or represented in different ways. For example, an event describing the visit of a webpage may appear in different ways in web browser logs and server logs.

In order to gather all the events found in footprint sources in a single timeline, a good handling of all these forms of heterogeneity is required. This leads to the development of an automated information processing approach which is able to extract knowledge from these heterogeneous sources. In addition, once extracted, this knowledge should be federated within the same model so as to facilitate their interpretation and future analysis. The effectiveness of a such approach can be assessed by the following criteria:

- Efficient automated tools that can extract events and build a timeline (Criterion 1 in Table 1).
- The ability to process multiple and various footprint sources and to federate the information collected in a coherent and structured way (Criterion 2 in Table 1).
- The ability to assist users during the timeline analysis. This latter encompasses many aspects such as making the timeline easier to read, identifying correlations between events or producing conclusions from knowledge contained in the timeline (Criterion 3 in Table 1).

For the majority of existing approaches, solutions are provided to automatically extract events and construct the timeline. Chen et al. (2003) introduced a set of automated extractors to collect events and store them in a canonical database, which allows to quickly generate a temporal ordered sequence of events. These automatic extractors, a widely used concept, can also generate the timeline (Olsson and Boldt, 2009; Gudhjonsson, 2010; Hargreaves and Patterson, 2012). However, current tools extract data in its raw form without a good understanding of the meaning of footprints, which makes their analysis more difficult. In order to deal with the semantic heterogeneity, the FORE (Forensics of Rich Events) system stores the events in an ontology (Schatz et al., 2004a,b). This ontology uses the notions of entity and event to represent the state change of an object over time. Nevertheless, the time model implemented in this ontology is not accurate enough (use of instant rather than interval) to represent events accurately. In addition, the semantic coverage of this ontology can be

Table 1
Evaluation of approaches.

Approach/Criterion	Auto extraction	Heterogeneity	Analysis	Theory	Data integrity
ECF (Chen et al., 2003)	✓	✓	✖	✖	✖
FORE (Schatz et al., 2004b)	✓	✓	●	✖	✖
Finite state machine(Gladyshev and Patel, 2004)	✖	●	●	✓	✖
Zeitline (Buchholz and Falk, 2005)	●	✓	✖	✖	✓
Neural networks(Khan and Wakeman, 2006)	●	●	✖	●	✖
CyberForensic TimeLab(Olsson and Boldt, 2009)	✓	✓	✖	✖	✖
log2timeline (Gudhjonsson, 2010)	✓	✓	✖	✖	✖
Timeline reconstruction (Hargreaves and Patterson, 2012)	✓	✓	●	●	✖