Knowledge Representation and Reasoning meets Digital Forensics: The COST Action DigForASP*

Stefania Costantini[1], Francesca Alessandra Lisi[2], and Raffaele Olivieri[1]

[1] Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica
Università degli Studi dell'Aquila, Italy
Stefania.Costantini@univaq.it,Raffaele.Olivieri@gmail.com
[2] Dipartimento di Informatica &
Centro Interdipartimentale di Logica e Applicazioni (CILA)
Università degli Studi di Bari "Aldo Moro", Italy
FrancescaAlessandra.Lisi@uniba.it

Abstract. Digital Forensics is a branch of criminalistics which deals with the identification, acquisition, preservation, analysis and presentation of the information content of digital devices. In this paper, we briefly describe DigForASP, a COST Action that aims to create a cooperation network for exploring the potential of the application of techniques from the field of Artificial Intelligence, in particular from the area of Knowledge Representation and Reasoning, in the Digital Forensics field, and to foster synergies between these fields. More precisely, in DigForASP the challenge is to address the so-called Evidence Analysis phase, where evidence about possible crimes and crimes' perpetrators must be exploited so as to reconstruct possible events, event sequences and scenarios related to a crime. Results from this phase are then made available to the involved stakeholders (law enforcement, investigators, public prosecutors, lawyers and judges). Reliability, explainability and verifiability of the results are therefore crucial.

Keywords: Knowledge Representation · Automated Reasoning · Digital Forensics.

1 Introduction

Digital Forensics (DF) is a branch of criminalistics which deals with the identification, acquisition, preservation, analysis and presentation of the information content of computer systems, or in general of digital devices, by means of specialized software, and according to specific regulations. In particular, the phase of *Evidence Analysis* involves examining and aggregating evidence about possible crimes and crime perpetrators collected from various electronic devices in

order to reconstruct events, event sequences and scenarios related to a crime. Evidence Analysis results are made available to law enforcement, investigators, intelligence agencies, public prosecutors, lawyers and judges.

Started in September 2018, the COST Action "Digital forensics: evidence analysis via intelligent systems and practices" (DigForASP)[3] aims at creating a research infrastructure for the application of *Artificial Intelligence* (AI), together with other complementary areas, in the field of Digital Forensics. DigForASP constitutes a timely challenge for both areas: DF and AI. From the AI perspective, the proposed research infrastructure will foster the development of new theoretical results, methods and techniques that will contribute in the long term to the development of new software tools that will rely on a complex combination of concepts and results from different areas of *Knowledge Representation* (KR) and *Automated Reasoning* (AR) such as diagnosis, causal explanation, temporal reasoning about actions, epistemic reasoning, the treatment of incomplete knowledge, deontic and legal reasoning, inductive learning and formal concept analysis, which will be complemented by other ones needed for the purpose of the Action. At the same time, the application of (intelligent) automated tools to DF - capable of reliable and exhaustive exploration of evidence, and with a level of analysis that goes beyond the scope of human observation and in time - will constitute a breakthrough that will have a direct impact on the practical investigation of crime scenarios.

To meet the challenge, the Action has built a network composed of researchers and engineers from the AI field together with DF experts belonging to Government Institutions and NGOs alongside scholars from the field of Information and Communication Technologies (ICT) and Law as well as social scientists, criminologists and philosophers (the latter for the ethical issues). The network is carrying out a set of activities and building resources to promote interaction, exchange and cooperation between these different areas. It is enabling computer scientists to understand the main issues and open problems of DF, especially Evidence Analysis, and it is helping to promote the exploitation of AI for addressing in an innovative, effective and adaptive way the key problems in this domain. Network partners is thus being able to identify KR&AR techniques which can be applied to Evidence Analysis, and to suggest guidelines for creating and developing suitable new techniques and methods aimed at advancing the state of the art in both DF and AI, strengthening European research and innovation capability in these areas. The long-term objective of the Network is to increase know-how and competences, so as to devise and to implement concrete projects and tools to be applied by Police Scientific Investigation Departments in solving real cases in so-called COST Member Countries, COST Near Neighbour Countries (NNCs) and COST International Partner Countries (IPCs).

The paper is organized as follows: In Section 2 we provide a short overview of the automated tools used in DF. In Section 3 we identify the challenges of DF to the areas of KR & AR. In Section 4 we briefly discuss the preliminary results achieved, and we conclude by outlining some promising directions for research.

[3] `https://digforasp.uca.es/`

2 State of the Art in Digital Forensics

Digital Forensics is a complex and rapidly evolving field, where methods for collecting evidence are varied, rapidly evolving and becoming increasingly sophisticated. Clearly, the development of DF is highly related to the development of ICTs in the last decades, and to the widespread diffusion of electronic devices and infrastructures. It involves various disciplines such as computer science, electronic engineering, various branches of law, investigation techniques and criminological sciences. Organizational aspects are also relevant and DF investigation involves, in general, several experts working with sophisticated instruments and software, with limited resources and tight timing.

To better understand the context of DF, we need to provide some details about how investigations are conducted. An investigation consists, in general terms, in a series of actions and initiatives implemented by the investigators (law enforcement and judges) in order to ascertain the "*truth*" and acquire all possible information and data about a perpetrated crime and related facts with their logical implications. A large number of entities are involved in this process, where they help to pursue a criminal activity, which could still be in progress. In an accurate vision, and according to the Italian Code of Criminal Procedure, investigations can be defined as "the set of activities carried out by the *officers* and *agents* of the *criminal police*". An investigation has therefore the ultimate goal of establishing the existence of a crime and the consequences that it has determined (generic proof or "*de delicto*"), and identifying the criminals (specific proof or "*de reo*"). These activities start from the acquisition of the crime notice or from the analysis of a crime scene. Through a series of initiatives and actions, the investigation allows the collection of data and elements which, according to certain deductive logical reasoning, should lead to draw conclusions. Investigative cases are usually complex, and involve a number of factors that need to be taken into account. Most of the collected data are nowadays obtained through digital devices and platforms either seized from the suspects, or available on the Internet or shared by telecommunication companies. This explains the increasing importance of DF in the legal process.

DF is divided into sub-fields according to the kind of data analyzed, including those extracted from the Internet, and it encompasses the following phases: (1) Identification, (2) Acquisition, (3) Preservation, (4) Evidence Analysis, and (5) Presentation. Phases 1-3 are supported by a number of hardware and software tools, the latter being both proprietary and open source. These tools are continuously evolving to follow the evolution of the involved technologies and devices, and recently related procedures have been standardized in all communities. However, they do not require advanced reasoning capabilities. Phase 4, i.e. Evidence Analysis, is where the main thrust of the Action will lie. It involves examining fragmented, incomplete knowledge, and aggregating evidence items into complex scenarios possibly involving time, uncertainty, causality and alternative possibilities. Currently, no single established procedure exists for Evidence Analysis, which is usually performed by Scientific Investigation experts on the basis of their experience and intuition. This phase requires therefore advanced

reasoning capabilities that are not currently supported by available devices and software. In fact, these are limited to data recovery (and data recognition) and to providing metadata (size, dates of creation/modification/elimination, etc.). Therefore, such retrieved data must be analysed by human experts, possibly with the support of available automated tools. However such tools, apart from text analysis, header files analysis and mining software packages, operate as a "black box" (i.e., they provide results without motivation or explanation), and for verification of the results one needs to perform a secondary analysis.

It should be acknowledged that AI techniques have been already applied to DF for different purposes. However, they have mainly focused on data retrieval and categorization tasks. The analysis of image and video files or the detection of anomalies in large databases such as email exchanges, network transactions, etc. are examples of such applications. These tasks benefit in particular from Machine Learning (ML) techniques. The Action takes a step beyond the state of the art as it fosters the application of KR&R methods to retrieved data in order to elicit evidence that can be used in a trial. For instance, from data items retrieved from different sources (like, e.g., mobile devices, social network activities, cloud computing tracks, etc.), we may obtain the set of all possible patterns of activity of a suspect during the execution of a crime. AR tools can constitute a crucial advantage since the amount of data to examine and interpret is large and keeps growing with the increasing adoption of digital devices in everyday life. Thus, the Action proposes innovations in the following two directions:

1. a substantial evolution of the current paradigm of evaluation and interpretation of data in DF, which might be exportable, in the future, also to other Forensic Sciences;
2. a breakthrough innovation for the judicial system, based on the possibility of adopting intelligent, reliable and dependable decision-support systems for the reconstruction of facts, able to take into account the wide number of elements and variables involved in complex cases, so as to aid judges in their assessments and decisions.

3 Research challenges for KR and AR in DigForASP

Unlike the phase of Identification, where the application of ML techniques can be useful for the analysis of big data, the phase of Evidence Analysis has particular requirements that make the proposal of DigForASP based upon KR and AR a much more promising approach, potentially becoming a breakthrough in the state-of-the-art. The ultimate goal of Evidence Analysis is the formulation of verifiable evidence that can be rationally presented in a trial. Under this perspective, the results provided by ML classifiers or other types of "black box" AI systems do not have more value than human witness' suspicions and cannot be used as legal evidence. Logical methods provide a broad range of proof-based reasoning functionalities that can be implemented in a declarative framework where the problem specification and the computational program are

closely aligned. This has the benefit that the correctness of the resulting systems can be formally verified. Moreover, recent research has led to new methods for visualising and explaining the results of computed answers (e.g., based on argumentation schemes). So one can not only represent and solve relevant problems, but also provide tools to explain the conclusions (and their proofs) in a transparent, comprehensible and justified way.

DigForASP aims at promoting formal and verifiable AI methods and techniques for Evidence Analysis. As already mentioned, this DF phase requires examining fragmented incomplete knowledge, and aggregating evidence items into complex scenarios. Relevant aspects to be considered include:

- Timing of events and actions;
- Possible causal correlations;
- Uncertainty and imprecision;
- Contexts in which suspicious actions occurred;
- Skills of the involved suspects;
- Awareness of the involved suspects of committing a violation or a crime and of the degree of severity of the violation/crime.

Moreover, given available evidence, several possible underlying scenarios may exist that should be identified, examined and evaluated. Currently, no single established procedure exists for this phase, which is usually performed by Scientific Investigation experts on the basis of their experience and intuition. The aim of the Action is that all the above should be performed via techniques that are verifiable with respect to the results they provide, how such results are generated, and how the results can be explained. Therefore, such software tools can be reliable and trustworthy, in the sense of confidence in the system's correct behaviour. Otherwise there remains an undesirable uncertainty about the outcome of these stages, and different technicians analyzing the same case can reach different conclusions which may lead to different judgements in court.

In AI, several methods and techniques have been developed over the years for uncertain, causal and temporal reasoning, and for devising and examining alternative consistent scenarios that might be compatible with a set of known facts. To the best of our knowledge, these techniques have never been applied to Evidence Analysis. Therefore, studying their applicability in this domain with the aim of developing suitable prototypes is per se a significant advance over the state of the art. Moreover, the application to such a challenging field will foster the refinement and/or improvement of known methods and techniques, and the development of novel ones. Overall, Evidence Analysis constitutes an ideal application domain for logical reasoning in AI, as it combines different classical aspects of KR&R. Also the two areas share a common feature: the search for a proof (a formal proof vs a valid argumentation in a trial). In the short term, DF can provide the AI community with non-trivial AR benchmarks that constitute a breakthrough with respect to available synthetic or ad hoc examples used in the scientific literature. It will act as a proof of concept to check whether different available KR&R techniques and tools are directly applicable or, most probably, require adjustments to take into account the domain features.

4 Achievements and Promising Directions

Modern investigative activities consist of well-established practical steps, such as the crime scene reconstruction, alibi verification, as well as the analysis of huge amounts of data coming from data files, smart-phone and telephone logs. So, as a first attempt, we have devised a reformulation of these activities which exploits provably correct encodings of known mathematical problems in order to elicit scenarios from DF data [8, 4]. In particular, we have considered to represent (fragments of) cases in Answer Set Programming (ASP), which is a well-established computational logic paradigm for dealing with computationally hard problems (cf., among many, [10, 11, 16, 1, 13, 9] for relevant literature). ASP has been selected for these preliminary experiments because of its ease of use and readability, for the availability of efficient freely available inference engines ("ASP solvers") and for the possibility of performing proof of correctness of the software (the reader may refer to [14, 12, 15] for the definition of the underlying formal properties). When applicable, the ASP encodings generate all possible scenarios compatible with the case's data and constraints. In the general case, this can be of great help as the human expert might sometimes overlook some of the possibilities: this has been verified by everyday practice, where different experts often generate different interpretations. Overall, leveraging the experience gained over the years by investigators, we have been able to claim with good reason that indeed a wide range of fragments of real cases can be mapped to computational problems, often to known ones.

In a future perspective, we may notice that logical methods (like ASP) could provide a broad range of proof-based reasoning functionalities (including, e.g., time and time intervals logic, causality, forms of induction, etc.) that can be integrated into a declarative framework for Evidence Analysis where the problem specification and the computational program are closely aligned. The encoding of cases via such tools would have the benefit that (at least in principle) the correctness of such declarative systems based on computational logic can be formally verified. Moreover, recent research has led to new methods for visualizing and explaining the results of computed answers (e.g., based on argumentation schemes). So, one could not only represent and solve relevant problems, but might also employ suitable tools to explain the conclusions (and their proofs) in a transparent, comprehensible and justified way. The engine of such a future Decision Support System might be based, again remaining within the computational logic realm, on Multi-Context Systems (MCS) [2, 3] and their agent-oriented extensions such as DACMACS (Data-Aware Commitment-based managed Multi-Agent-Context Systems, [6, 7]) and ACEs (Agent Computational Environments, [5]). Agent-based simulation is also a promising direction to explore in Evidence Analysis, since it allows the generation of plausible scenarios. Finally, besides deduction, other inferences should be considered, notably induction and abduction.

Acknowledgments The network "DigForASP - Digital forensics: evidence analysis via intelligent systems and practices" is funded by the European Cooperation in Science and Technology (COST) under the Horizon 2020 framework programme.

References

1. Baral, C.: Knowledge representation, reasoning and declarative problem solving. Cambridge University Press (2003)
2. Brewka, G., Eiter, T., Fink, M., Weinzierl, A.: Managed multi-context systems. In: IJCAI 2011, Proceedings of the 22nd Intl. Joint Conf. on Artificial Intelligence. IJCAI/AAAI (2011)
3. Brewka, G., Ellmauthaler, S., Pührer, J.: Multi-context systems for reactive reasoning in dynamic environments. In: ECAI 2014, 21st European Conf. on Artificial Intelligence, Proceedings. IJCAI/AAAI (2014)
4. Costantini, S. De Gasperis, G. and Olivieri, R.: Digital forensics and investigations meet artificial intelligence, Annals of Mathematics and Artificial Intelligence, in press (2019)
5. Costantini, S.: Ace: a flexible environment for complex event processing in logical agents. In: Engineering Multi-Agent Systems, Third International Workshop, EMAS 2015, Revised Selected Papers. Lecture Notes in Computer Science, vol. 9318. Springer (2015)
6. Costantini, S.: Knowledge acquisition via non-monotonic reasoning in distributed heterogeneous environments. In: 13th Int. Conf. on Logic Programming and Nonmonotonic Reasoning LPNMR 2013, Proceedings. Lecture Notes in Computer Science, vol. 9345. Springer (2015)
7. Costantini, S., De Gasperis, G.: Exchanging data and ontological definitions in multi-agent-contexts systems. In: Paschke, A., Fodor, P., Giurca, A., Kliegr, T. (eds.) RuleMLChallenge track, Proceedings. CEUR Workshop Proceedings, CEUR-WS.org (2015)
8. Costantini, S., Olivieri, R.: Digital forensics evidence analysis: An answer set programming approach for generating investigation hypotheses. In: 13th Int. Conf. on Logic Programming and Nonmonotonic Reasoning LPNMR 2015, Proceedings. Lecture Notes in Computer Science, vol. 9345, pp. 242–249. Springer (2015). Presented also at CILC 2015, 30th Italian Conference of Computational Logic, CEUR Workshop Proceedings 1459, CEUR-WS.org
9. Erdem, E., Gelfond, M., Leone, N.: Applications of answer set programming. AI Magazine **37**(3), 53–68 (2016)
10. Gelfond, M., Lifschitz, V.: Classical negation in logic programs and disjunctive databases. New Generation Computing **9** (1991)
11. Leone, N.: Logic programming and nonmonotonic reasoning: From theory to systems and applications. In: Logic Programming and Nonmonotonic Reasoning, 9th Intl. Conference, LPNMR 2007, Proceedings. Springer (2007)
12. Lifschitz, V., Pearce, D., Valverde, A.: Strongly equivalent logic programs. ACM Transactions on Computational Logic **2**, 526–541 (2001)
13. Lifschitz, V.: Twelve definitions of a stable model. In: Proceedings of the 24th Intl. Conference on Logic Programming. Lecture Notes in Computer Science, vol. 5366, Springer (2008)
14. Pearce, D.: A new logical characterization of stable models and answer sets. In: Non-Monotonic Extensions of Logic Programming, pp. 55–70. Lecture Notes in Artificial Intelligence, vol. 1216, Springer (1997)
15. Pearce, D., Valverde, A.: Synonymous theories in answer set programming and equilibrium logic. ECAI 2004, 16th European Conf. on Artificial Intelligence, Proceedings. IJCAI/AAAI (2004)
16. Truszczyński, M.: Logic programming for knowledge representation. In: Logic Programming, 23rd Intl. Conference ICLP 2007, Proceedings. Springer (2007)