Funded by the European Commission
Seventh Framework Programme

Cyber ROAD

Development of the Cybercrime and Cyber-terrorism Research Roadmap

Grant Agreement N. **607642**

Forensics

Author(s): Peter Kieseberg (CDF), Davide Andreoletti (SUPSI), Angelo Consoli (SUPSI)

CT RG-58 and CC RG-62 – There is a huge research gap regarding database forensics in general, especially considering the detection of manipulation by users with higher privileges and even database administrators.
CT RG-71 – Intelligence has access to great volume of information but lacks of tools to identify the most meaningful
CT RG-72 – Research on cyberterrorism motivations/root cause
CT RG-79 – Improve Incident Response
CT RG-93 and CC RG- 98– Methods for forensics and crosschecking. Methods for assessing the sensitivity of data aggregates and linked data sets.
CC RG- 101 and CC RG-103 – Watermarks and Fingerprints that cannot be spotted easily and is resilient to collusion attacks and inferences.

ABSTRACT

Digital Forensics is often solely associated with the areas of attribution in the aftermath of a cyber-attack, as well as the discovery of hidden and the restoration of deleted files, either due to malicious attempts or simply because of errors by a legitimate user.

Still, digital forensics possesses many aspects that are not covered by current large-scale research and could be useful in order to thwart acts of cyberterrorism and cybercrime. Efficient methods for enabling cheap and fast digital investigations could not only lead to a much faster attribution of attacks with subsequent countermeasures (reactive, as well as proactive with respect to follow-up attacks), which would in itself be a valuable issue in order to reduce the expected motivation and increase the danger for cybercriminals. Accessible forensic methods could also be used in order to even detect attacks, especially considering instances of data and system manipulation. With respect to data theft and industrial espionage, methods for data leak detection and comprehensive audit & control mechanisms could help reduce the danger of espionage and exfiltration attacks by making the data thief more easily detectable.

Digital forensics and its methods is always considered to be a mixed blessing, blurring the line between rightful and lawful investigation and the protection of user rights. This is especially important in case of comprehensive auditing of access patterns and employee work: Here, research into the harmonization of European law will be needed, allowing for solutions that provide a certain level of detection of data theft and manipulation, as well as protecting the privacy of the innocent. While this legal work is of course of great importance, it needs a strong relation to technological research in the area of digital forensics, in order to make laws that can be related to the technological reality.

With the advent of cheap sensors that allow the wide adoption of new paradigms like IoT, new challenges for forensic methods have appeared: Opportunistic networks and ad hoc connections increase the difficulty of forensic investigation drastically. Still, new technologies in the area of machine learning and data mining will also open up new doors with respect to real-time forensics (live forensics) based on patterns and behaviour. Furthermore, the introduction of principles like industry 4.0 and personalized/participatory medicine introduced completely new classes of equipment and environments that need new solutions: For example, industry lanes might not be capable of shutting down several days for an investigation as it is currently done in the more traditional IT-world, without causing huge losses.

Summarized, the introduction of new technologies poses a large amount of new challenges to the field of digital forensics, which need further development in order to answer these challenges.

Forensic methods for Big Data

DISTANCE TO THE MARKET: **TRL 2**
COST OF THE TOPIC: **3 STREPs + 2 IP**
AVAILABILITY OF COMPETENCE IN EUROPE: **5**
TIME SPAN FOR ADDRESSING THE ACTION: **36 months**
ACTORS: **Big Data providers, LEAs, Machine Learning specialists, Security experts**

Big Data is one of the major emerging topics of the last years, especially considering the multitude of related issues: Big sensor networks in the IoT-world that generate large amounts of data in real time with complex linkage, as well as data driven science and economy, ranging from biomedical research in the area of personalized medicine based on genomics to providing added services in large industrial environments within the new industry 4.0 paradigm.

Manipulations of such big data streams are currently very hard to detect, especially in case of unspecified sources. Here, machine learning algorithms can help to detect changes in the delivered data, or irregularities in the data provisioning.

Another big challenge for data driven environments with respect to Privacy is the question on detecting (singular) sensitive data particles within the large amount of data. This is especially important in environments where the data is not derived by one or a few well-defined sources, but rather form an agglomeration of data provided by different sources with variable content. In case of overlapping source information, this can pose even more important questions, as most anonymization strategies in use are susceptible to de-anonymization through collusion attacks, i.e. the anonymization can often be endangered by aggregating different versions of the same data set. Thus, in order to pursue the goal of guaranteeing the protection of sensitive data, we propose to use machine learning techniques that are able to 1) detect sensitive information in large data streams and 2) rate the combination of different data sources with respect to linkage possibilities.

Live Forensics, Audit & Control

DISTANCE TO THE MARKET: **TRL 7**
COST OF THE TOPIC: **2 STREPs + 2 IP**
AVAILABILITY OF COMPETENCE IN EUROPE: **5**
TIME SPAN FOR ADDRESSING THE ACTION: **32 months**
ACTORS: **industry, IoT-experts, Security experts**

Traditional forensic investigations are set up as ex-post analysis in post-mortem or post-incident investigations, often based on a detected or assumed irregularity in the system. This approach comes with some deficits with respect to the new environments introduced within the industry 4.0 paradigm or large IoT-structures:

1. The analysis happens ex ante, i.e. the damage is already done. Modern environments need the detection of attacks and manipulations to be much faster, in order to avert damage. A possible improvement is the development of methods for generating a complete and reliable overview on the health status of each endpoint in the system.
2. Typically, the system under investigation is either shut down for normal operation, or an exact copy of the system is generated, where the analysis is then conducted on. Both approaches are completely unfeasible in more complex production systems, as a shutdown of the system typically results in large financial damages, in case of some applications like steel furnaces, even in a complete shutdown of the factory. On the other hand, since large IoT or industrial systems are extremely expensive and highly proprietary in nature, the production of an exact copy is not feasible.

Based on these problems, forensic investigations need to enable infrastructure owners to efficiently monitor their systems and conduct forensic investigations directly on the running system.

Thus, the results of this research action shall include:

- The development of modern audit & control applications that monitor the overall health status of systems, including access to resources and system components. This also includes the development of working solutions that can be introduced into standard of the shelf IoT-structures.
- The development of tools and processes for introducing live forensics into industrial environments, i.e. enabling the forensic analyst to analyse complex systems without shutting them down: critical information needs to be protected in order to be used during the analysis, while the investigation process must not interfere with the system stability. This will also include methods for efficiently generating forensic clones of industry systems.

Forensics in mobile and distributed Environments

DISTANCE TO THE MARKET: **TRL 5**
COST OF THE TOPIC: **4 STREPs + 2 IP**
AVAILABILITY OF COMPETENCE IN EUROPE: **5**
TIME SPAN FOR ADDRESSING THE ACTION: **36 months**
ACTORS: **mobile services providers, Industry, IoT-Provider**

With the rise of smartphones, mobile platforms gained increasing popularity, which also had drastic effects on the area of forensics: nowadays, these devices are one of the main targets during forensic investigations, often in course of analysing traditional scenarios of crime and terrorism (see also the discussion regarding decryption of iOS-based devices by the FBI).

Still, there are many technological issues to overcome, especially with respect to the new device-side encryption technologies that are currently increasingly implemented. While the smartphone arena is receiving increasing attention, the widespread implementation of IoT-devices like integrated sensors relates to new issues in the area of forensics: due to some features characterizing opportunist ad-hoc networks (e.g., energy saving and mobility of the sensors) traditional methods for network forensics, as well as network observation, are not suitable anymore. Thus, a key step of future research concerns the development of new techniques that allow to analyse the network structure at the exact time when the attack was carried out, even during an ex-post analysis, including monitoring features and the possibility to gather all network changes.

One of the main issue in mobile forensics lies in the area of legal exploitability: while many techniques exist that can be used in order to gain forensic information on system manipulation and disruption, close to none of them hold in front of court. Most often the problem lies in the complexity of the forensic tools and methods at hand that make the trustworthiness of the gathered results hard to determine, i.e. whether the evidence was not fabricated by the forensic investigators. Thus, it is of paramount importance to develop technologies, processes and especially tools that fulfil all the legal requirements in order to be acceptable as proof in front of court.

Following these principles as much as possible during forensics will diminish incidents regarding the authenticity of evidence, and augment the transparency of the whole forensic process.

RESEARCH ACTION #4

Database Forensics and Data Leak Detection

DISTANCE TO THE MARKET: **TRL 5**
COST OF THE TOPIC: **4 STREPs + 2 IP**
AVAILABILITY OF COMPETENCE IN EUROPE: **5**
TIME SPAN FOR ADDRESSING THE ACTION: **36 months**
ACTORS: **Security researchers, data driven domains, industry 4.0, big data analysts**

The large data dependent systems that are currently in use everywhere are unimaginable without the use of databases. Most of the data stored in all kinds of applications and systems is stored in some kind of database management system (DBMS). Furthermore, data has become a resource and thus a (monetary) asset in many modern, often service driven, environments: starting from research labs in the bionics sector up to data driven industrial services, data nowadays not only needs protection due to privacy concerns, but simply as a valuable resource.

Data manipulation and data theft are therefore important attack vectors for cyberterrorists and cybercriminals respectively. While for data stores on file systems bases there exists a multitude of different approaches for manipulation detection, file recovery and other techniques typically utilized during a forensic investigation, the field of database forensics has been left rather unattended during the last decades. Furthermore, most recent works solely concentrate on attacks carried out by "normal" users or even outsiders, missing attacks that are carried out by an administrator, either due to a malicious insider, or in the course of stolen credentials through either a previous technical attack or social engineering. These attacks are far more dangerous, since database administrators typically have the possibilities to conceal their tracks (e.g. in log files and audit mechanisms) effectively and can stay undetected for a very long time. Another issues lies in the release of data sets in order to conduct joint projects with other (often industrial) partners. As this data is only shared for a certain cause with a strictly limited amount of partners, it must not be passed on. While passing on might not be an issue with respect to privacy, even insensitive data (e.g. sensor measurements) can possess a huge value. Thus, methods that allow to detect leaking parties even in case of incomplete data sets are needed. Current approaches typically rely on the introduction of marker data that either introduces slight errors in case of analysis, which can cause problems in case of data driven analysis, or rely on finding large portions of the leaked set. The aim of this action is to develop techniques for quick and unambiguous detection of data leaks based on single records or attribute subsets.