Extracting Intelligence from Digital Forensic Artefacts

Dr. Stilianos Vidalis[1], Dr Olga Angelopoulou[1], Prof. Andy Jones[2]
[1]Cyber Security Centre,
School of Computer Science,
University of Hertfordshire, UK
[2]Security Research Institute,
Edith Cowan University, Australia
s.vidalis@herts.ac.uk
o.angelopoulou@herts.ac.uk
a.jones26@herts.ac.uk

Abstract: Forensic science and in particular digital forensics as a business process has predominantly been focusing on generating evidence for court proceedings. It is argued that in today's socially-driven, knowledge-centric, virtual-computing era, this is not resource effective. In past cases it has been discovered retrospectively that the necessary information for a successful identification and extraction of evidence was previously available in a database or within previously analysed files. Such evidence could have been proactively used in order to solve a particular case, a number of linked cases or to better understand the criminal activity as a whole. This paper will present a conceptual architecture for a distributed system that will allow forensic analysts to forensically fuse and semantically analyse digital evidence for the extraction of intelligence that could lead to the accumulation of knowledge necessary for a successful prosecution.

Keywords: intelligence-led policing, evidence fusion and dissemination, forensic intelligence, ID theft

Setting the scene

A few years ago, a case was brought to a successful conclusion. The suspect was convicted for a number of computer-related crimes based on digital evidence that were extracted from computing devices found in his possession, following the standard and very well published dead-box digital forensics analytical procedure. After the end of the proceedings, one of the authors was given access to the evidence for further analysis. Such analysis was previously considered outside the scope of the investigation. The author was then able to extract actionable intelligence, linking the convict to a more serious crime and a number of other criminal activities on a different continent, committed in collaboration with a number of foreign nationals.

In the past, computer-related crimes were defined as those activities where computers were used for the commission of crime, where computers contained evidence of crime and/or where computers were the targets of crime (Hale 2002). Given today's (socially-driven knowledge-centric virtual-computing era) specific parameters, there is a need for a slightly different and more inclusive definition, addressing the different types of computing devices, e.g. mobile phones, smart embedded devices, game consoles, laptops, computers, etc. and a domain that goes beyond the concept of the term "cyber-domain". For the purposes of our research we will use the term Information Environment (IE). The U.S. Department of Defence (DOD) has defined the Information Environment (IE) in its Joint Publication 3-13 for

Information Operations (US DoD 2012), stating that *"... the information environment is the aggregate of individuals, organizations and systems (resources) that collect, process, disseminate, or act on information."* Hence, computer-related crimes can be defined as:

> *Activities where physical and logical computing devices, attached to an Information Environment, are used for the commission of crime, contain evidence of crime, and/or are the targets of crime.*

Continuing our reasoning, and in agreement with the Association of Chief Police Officers (ACPO) guidelines, digital evidence, or computer-based electronic evidence is information and data of investigative value that is stored on or transmitted by a computer. Menou (1995, as cited in Chowdhury and Vidalis, 2013) described information as "*a product, which encompasses information as thing, as object, as resource, as commodity, what is carried in a channel (including the channel itself), the contents*."

Combining all of the above definitions, and accepting that information is paramount to the resolution of any crime, we can also accept that only having forensic evidence is no longer adequate for resolving computer-related criminal activities. Today, investigators need to have forensic intelligence (Ribaux et.al. 2003), (Legrand and Vogel 2012), even for the simplest and most trivial computer-related crime, that can lead to forensic evidence which, when combined, can lead to a strong supporting case for a prosecution. Such intelligence can be used either in a pro-active or in a re-active manner. As a concept, this is not new. It was first introduced and discussed a number of decades ago (Birkett 1989), (Ribaux and Margot 1999). For example, in the UK, ENDORSE (National Crime Agency 2015) is a nation-wide forensic and law enforcement initiative to collect and analyse information from drug seizures made in the UK. Apropos, the use case for ENDORSE is limited to a specific problem and a specific crime type within one national jurisdiction. Furthermore, computer-related criminal activities can be seen as a very complex problem, combining different types of traditional criminal activities with different and innovative technologies for transcending jurisdictional boundaries.

Intelligence is the timely, accurate and usable product extracted from logically processed information. For this extraction to be successful and accurate, one must apply specialist knowledge. In the case of forensic intelligence, the concept of knowledge is twofold:

a. procedural knowledge on the identification, individualisation, association and reconstruction of forensic evidence, and
b. crime-specific analytical knowledge for translating leads to actionable forensic intelligence.

Before discussing the requirements for a system able to handle forensic intelligence, it is considered beneficial to identify problems with the current practice of resolving computer-related criminal activities.

Operational level issues

It is indicated (Statistic Brain 2015), (Mkomo 2015) that the cost of storage per GB of data is currently $0.03. Cloud storage is even cheaper according to The Register (2014). The cost of undertaking criminal activities is coming down. Even worse, because of the latest computing innovations such as virtualisation, cloud applications, communication applications and connectivity and interconnectivity opportunities, the physical crime scene is often different from the logical cybercrime scene.

According to Statista.com, 364.59 million Hard Disk Drives (HDDs) shipped globally in the first three quarters of 2015, and a figure of 416.7 million HDDs and 153.8 million Solid State Drives (SSDs) was projected for the whole of 2015. According to Digitaltends.com, the average size of the Seagate HDDs is now over one terabyte. Based on these statistics, we can assume that a typical case would require Law Enforcement Agency (LEA) Officers to collect, on average, more than 1TB of data (including CDs, DVDs, internal and external HDDs/SSDs). The automated procedures that can be used to assist in the processing of this data, such as file signature analysis and hash analysis, are employed. Apropos, a large amount of data has to be manually analysed. Even before the analysis stage, there is a lot of work to be undertaken. As part of a testing activity in a digital forensics laboratory, the authors had to clean a hard disk. Forensically wiping one Samsung HD105SI 1TB drive, using a tableau TD2u, was achieving an average of a 6.6GB/min transfer rate and a projected turnaround time of 2h 30 minutes. Furthermore, in a recent disk study the authors performed, a large number of hard disks were acquired and forensically analysed. The average acquisition transfer rate that was achieved was 2.76GB/min. This translates on an average time investigators would need to spend in the acquisition phase of at least 6 hours per disk.

After the acquisition of the devices, a forensic analyst will get to the analysis phase, where, depending on the case, they will perform any/all of the following activities:

- Disk geometry analysis (number, size and type of partitions (deleted or not))
- Time-zone analysis
- Operating System analysis
- Hash analysis
- File signature analysis
- Registry analysis
- Compound file analysis
- Log file analysis
- Internet artefacts analysis
- Email analysis

Following the above, more specific analytical steps will have to be performed (the list is not meant to be comprehensive):

- Deleted files recovery
- Identification of USB devices that were ever connected and when they were connected
- Identification of files and folders that have been exfiltrated
- CD/DVDS that may have been burned
- Websites visited and by which user account
- Lists of recently used programs, the files they have accessed, and when they have done so
- Programs that have been installed and uninstalled
- Attempts at data destruction/hiding
- Program settings that can deduce knowledge of an act or technology

- What programs start when the computer starts and any related DLLs, cross-examining findings for the identification of malware footprints
- How many times a program has ever been run and by which user account
- Wi-Fi connection points that have been accessed and when
- Hidden email and other internet accounts
- Identify and analyse photos and deduce the geographic location of where photos were taken
- What particular user performed a task (related to the above activities or to case specific activities)

Nowadays, most of the above analytical tasks have been automated. Still, depending on the datasets used, the analysis phase will take on average at least two days per disk to complete. This translates in two days per disk before the forensic analyst will be able to start the manual analytical activities, the file indexing and any case-specific raw searches. It also translates in two days that physical computing resources will have to be locked down and assigned to the execution of the aforementioned tasks.

Operationally, all of the above issues have a significant impact and create backlogs that, over time, can become unmanageable. An Open Source Intelligence (OSINT) search conducted in February 2015 by the authors indicated that police forces in the UK have backlogs that range from anywhere between 12 to 24 months. To overcome this issue, non-specialist staff are deployed, using of-the-self triaging products to minimise the collected artefacts, and even then, often the evidence is split between different analysts, with different levels of experience, in different teams, following slightly different analytical SOPs. Even worse, often LEA analysts stop their analysis once they have identified/extracted enough evidence to support an argument for a successful prosecution.

Our hypothesis is that analysts can make use of forensic intelligence, which once combined with forensic evidence can streamline and optimise the analytical process in a cost effective and appropriate (from a penology perspective) manner.

Requirements for fusing and semantically analysing evidence

In order to support our hypothesis, we will use identity theft as an example cyber-crime type. Identity theft is one of the major concerns today (CIFAS 2015), (ITRC 2015), (Harell 2015). It has a significant human component and is being strongly influenced by the way people treat personal information (defined and discussed in the doctoral thesis of Dr. Angelopoulou). BBC, amongst other reputable online sources, has published that in the first quarter of 2015 the number of ID-theft victims rose by 31% (BBC 2015). These statistics suggest that it is extremely difficult to eliminate identity theft by employing stronger computer security techniques since the perpetrators constantly find ways around them. Furthermore, based a report from Experian (2015) individuals often do not adopt any measures to protect themselves online. The average person easily provides and shares personal identity information relating to one or more of their online aliases.

> *"Regard your good name as the richest jewel you can possibly be possessed of - for credit is like fire; when once you have kindled it you may easily preserve it, but if you once extinguish it, you will find it an arduous task to rekindle it*

again. The way to gain a good reputation is to endeavour to be what you desire to appear". Socrates (469BC-399BC)

Personal identity information (PII) is increasingly being stored and used in a range of digital forms. If this information is not adequately protected, this can leave individuals exposed to a range of possible threats. Identity Theft (ID theft) is defined as someone's action of using any sort of distinct personal private information with fraudulent intention; mainly for financial gain (Angelopoulou et.al 2007). Technology related examples include; identity theft malware and key loggers, phishing, web-spoofing, online social engineering and database data retrieval. The complexity of retrieving substantial evidential information of an ID theft crime perpetrated over the Internet demands a specific methodological instrument that will be able to identify and extract evidential components (forensic intelligence) from different cases (in a big data analysis context). When such incidents are examined in detail, then the nature of the problem can be more clearly understood.

For our example we are utilising a platform that has the ability to extract, fuse and share ID-theft related forensic intelligence. The Standardised Forensic Intelligence Platform (SFIP), which is described in the following sections, runs the Identity Theft Investigations (ITI) module to extract, fuse and share ID-theft related forensic intelligence. The intelligence data will be hosted on a unified database from which LEA Officers will be able to collect case related knowledge relating to the suspect, the victim and the modus operandi MO, while waiting for the digital forensic analytical activities to conclude. The use case for our proposed system is presented in figure 1.

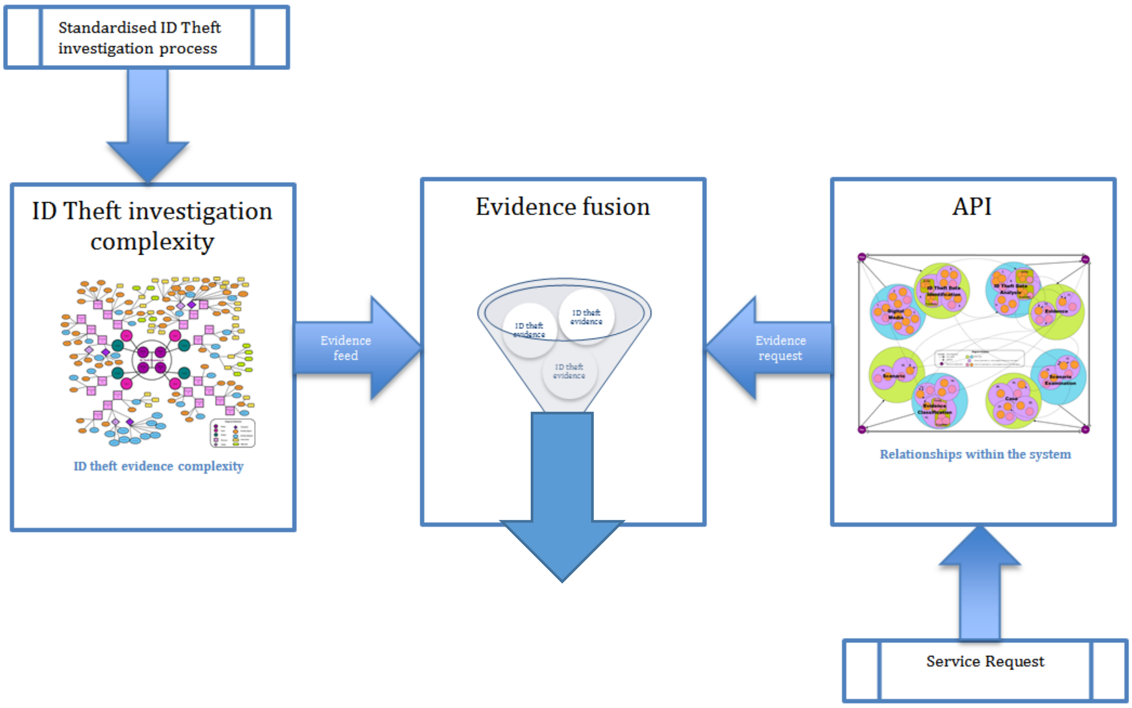

Figure 1: SFIP use case

The intelligence, knowledge and expertise will result from classified evidence that will be produced after a certain process and analysis has been followed. The fusion and semantic analysis of the evidence should provide a systematic application to a digital investigation of a cybercrime. We have summarised the requirements such a system should adhere to in the following list:

- Locally deployed modules must be connected to the local analytical workstations running different toolkits from different vendors. There is a requirement for a plug-in mechanism to allow for the development of new application program interfaces (APIs) as new forensic toolkits come to the market.
- The local database holding the intelligence data must replicate the manner in which data are being recorded by the Forces and obviously comply with ISO2700.
- Nodes must communicate in a secure manner. Encryption of all communication channels and of all messages (multi-layered encryption).
- Authentication and validation of all nodes and users.
- Signatures for providing non-repudiation and message integrity for all communications between system stakeholders/users/actors
- Logging of user and system requests and activities to ensure and assure chain of custody
- A metadata distribution system to ensure and assure evidential integrity and validity/authenticity of assets/artefacts.

Conceptual architecture

The proposed system will guide the forensic analyst through the intelligence collection process as the standard operating procedure (SOP) steps/activities will be built into the system. In our example as illustrated in figures 1 and 2, we are using the ITI module and an id-theft crime specific SOP. Semantic analysis using crime specific ontologies will be employed in order to extract crime-specific intelligence. Semantic techniques will enable contextual and relevant data to be identified for a particular entity. The use of ontologies will create a bridging mechanism, whereby semantic metadata could be referenced and validated to ensure that relevant and useful information is collected. This also ensures that trust and logic can be attained in the service request functionality. In the ITI module for example, the extracted knowledge (from the intelligence data) will link devices and content from different cases together, providing investigators with leads and forensic evidence that can be used for the profiling and prosecution of perpetrators.

SFIP will effectively present next to real-time knowledgeable answers to runtime user generated queries. It will collect information from disparate sources and use semantics to safeguard the future of knowledge discovery and reuse. The stakeholders will be able to request access to SFIP through an application program interface (API) and query specific relationships within the system.

The conceptual architecture of the system is illustrated in figure 2.

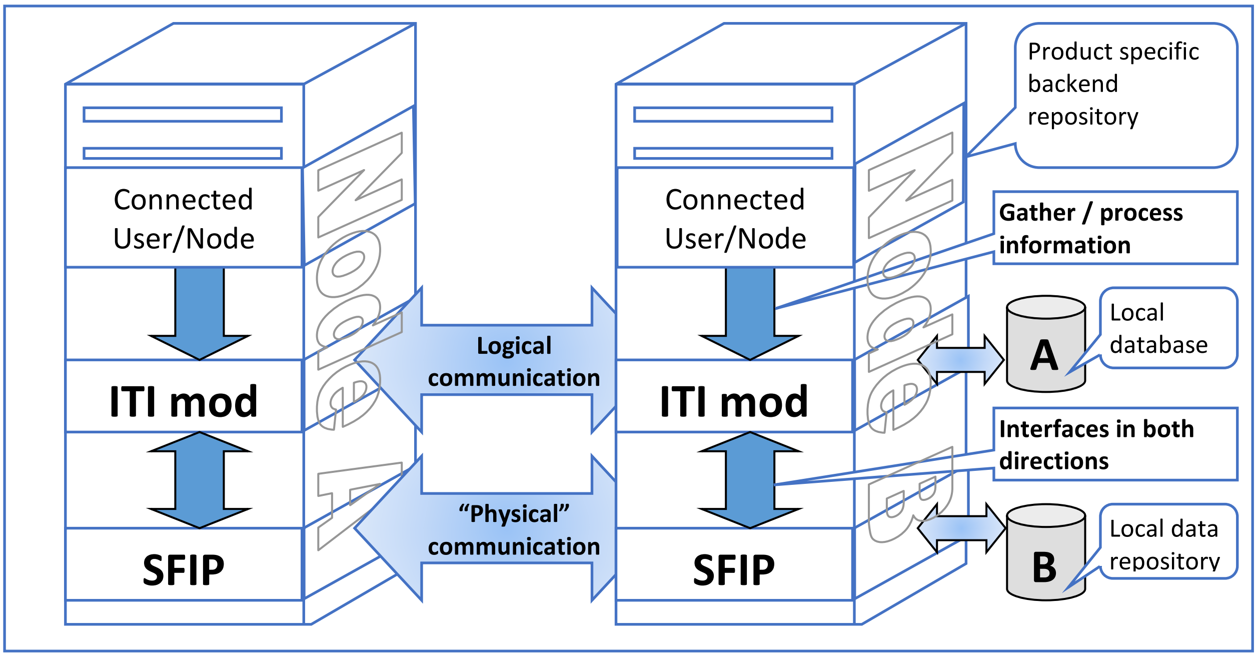

Figure 2: Conceptual architecture (original in Pilgermann et al. 2005)

The exchange of information is comparable with the ISO 7-layer OSI model (Stevens and Wright 1995), (Pilgermann and Blyth 2004). Although, a logical connection is established between the crime-specific modules of the nodes, they are not able to communicate with each other directly. Instead, they are making use of communication facilities provided by the SFIP layer. Furthermore, the crime-specific modules gather and process information from connected analytical workstations. Each node in the overall topology may act as both a source and a consumer of intelligence. However, regarding to roles, certain nodes may only be allowed to either send information or receive information. Each SFIP node maintains its database for storing all information about current and past cases (marked as 'A' in Figure 2). The employed technologies in each layer will be directly addressing the issues of security (both the channel and the content), authentication of nodes and users, and non-repudiation of transactions between the layers and the nodes. Generated logs will be managed as evidence, applying best practice on evidence handling and management as specified by ACPO.

Conclusion

Fusing forensic evidence from different cases, performing 'big-data-like' analytical activities, extracting forensic intelligence regarding persons (perpetrators and victims), assets and MOs, is believed to be the future of resolving computer-crime related activities. Law enforcement strategic vision certainly reflects the above as currently (first quarter of 2016) there is an invitation for tenders for a cybersecurity digital service infrastructure. The aim of the European call is to launch a core service platform that will serve national and/or governmental CSIRTs and CERT-EU. The harmonisation of collaboration procedures between CSIRTs is indeed envisaged to improve cooperation between them and equip them for a better handling of threats to cyber resilience in the European Union.

The proposed system directly addresses the identified requirements in the invitation for tenders. The authors are currently setting up a test-bed that will allow for the prototype development of the proposed system and of the crime-specific analytical modules. The authors are also establishing user groups for the creation and sharing of datasets that will be used for evaluation purposes. When the test-bed goes operational, statistics on the turnaround time of analytical tasks will be generated in order to understand if the proposed system can indeed provide LEAs with an immediate solution to the shortfalls of their current practice. At a later stage, a second data-set on the identified forensic intelligence will be created in order to understand and appreciate the conversion of intelligence enriched cases to successful prosecutions.

References

Hale, C. (2002) “Cybercrime: Facts & figures concerning this global dilemma”, Crime and Justice International, Volume 18, Issue 65

U.S. Department of Defense. (2012) Information operations. Joint Publication: 3-13. Retrieved from http://www.Dtic.Mil/Doctrine/New_Pubs/Jp3_13.Pdf

Chowdhury, T., and Vidalis, S. (2013) Proactively defending computing infrastructures through the implementation of live forensics and website capture in corporate network security. Third International Conference on Cybercrime, Security and Digital Forensics, Cardiff University, Cardiff, UK, June.

Ribaux, O,. Girod, A., Walsh, S,J,. Margot, P,. Mizrahi, S,. Clivaz, V,. (2003) “Forensic intelligence and crime analysis”, Law Probability and Risk, Volume 2, issue 1, pp 47-60.

Legrand, T., and Vogel, L., (2012) Forensic Intelligence, https://www.academia.edu/1519407/Forensic_Intelligence

Birke, J. (1989) Scientific scene linking, Journal of the Forensic Science Society, volume 29, Issue 4, pp 271-284

Ribaux, O., Margot, P., (1999) “Inference structures for crime analysis and intelligence using forensic science data: the example of burglary”, Forensic Science International, Volume 100, Issue 3, pp 193-210

National Crime Agency, (2015) Forensic Intelligence, http://www.nationalcrimeagency.gov.uk/crime-threats/drugs/forensic-intelligence

Statistic Brain, (2015) http://www.statisticbrain.com/average-cost-of-hard-drive-storage/)

Mkomo, (2015) http://www.mkomo.com/cost-per-gigabyte-update)

The Register, (2014) http://www.theregister.co.uk/2014/03/25/google_price_slash/

CIFAS, (2015) Fraudscape UK fraud trends, United Kingdom, http://www.cifas.org.uk/secure/contentPORT/uploads/documents/External%20-%20Fraudscape%20main%20report%20for%20website.pdf

ITRC, (2015) Identity Theft Resource Center Breach Report Hits Record High in 2014, *Identity Theft Resource Center*, http://www.idtheftcenter.org/ITRC-Surveys-Studies/2014databreaches.html

Harrell, E., (2015) *Victims of Identity Theft, 2014*, U.S. Department of Justice, http://www.bjs.gov/content/pub/pdf/vit14.pdf

BBC, (2015) Number of identity theft victims rises by a third, http://www.bbc.co.uk/news/uk-32890979

Experian, (2015) One in six adults has fallen victim to cyber-crime, http://www.experian.co.uk/blogs/latest-thinking/one-six-adults-fallen-victim-cyber-crime/

Angelopoulou, O., Thomas, P., Xynos, K., Tryfonas, T., (2007) Online ID-Theft techniques, investigation and response, International Journal of Electronic Security and Digital Forensics, Volume 1, Issue 1, pp 76-88

Pilgermann, M., Vidalis, S., Blyth, A., (2005) "Inter-Organisational Intrusion Detection Using Knowledge Grid Technology", Journal of Information Management and Computer Security, Volume 14 Number 4.

Stevens, W. R. and Wright G.R., (1995) TCP/IP Illustrated, Volume 2 - The Implementation. USA, Addison-Wesley Publishing Company

Pilgermann, M. and Blyth, A., (2004). “Anonymizing Data in a Peer-To-Peer based Distributed Intrusion Detection System - A possible Approach” European Conference on Information Warfare (ECIW), London.