12-4-2013

Towards An Automated Forensic Examiner (AFE) Based Upon Criminal Profiling & Artificial Intelligence

M Al Fahdi
Plymouth University

N L. Clarke
Plymouth University

S M. Furnell
Plymouth University

DOI: 10.4225/75/57b3be61fb866
11th Australian Digital Forensics Conference. Held on the 2nd-4th December, 2013 at Edith Cowan University, Perth, Western Australia

TOWARDS AN AUTOMATED FORENSIC EXAMINER (AFE) BASED UPON CRIMINAL PROFILING & ARTIFICIAL INTELLIGENCE

M. Al Fahdi, N.L. Clarke & S.M. Furnell
Centre for Security, Communications & Network Research (CSCAN)
Plymouth University, Plymouth, United Kingdom
info@cscan.org

Abstract
Digital forensics plays an increasingly important role within society as the approach to the identification of criminal and cybercriminal activities. It is however widely known that a combination of the time taken to undertake a forensic investigation, the volume of data to be analysed and the number of cases to be processed are all significantly increasing resulting in an ever-growing backlog of investigations and mounting costs. Automation approaches have already been widely adopted within digital forensic processes to speed up the identification of relevant evidence – hashing for notable files, file signature analysis and data carving to name a few. However, to date, little research has been undertaken in identifying how more advanced techniques could be applied to perform "intelligent" processing of cases. This paper proposes one such approach, the Automated Forensic Examiner (AFE) that seeks to apply artificial intelligence to the problem of sorting and identifying relevant artefacts. The proposed approach utilises a number of techniques, including a technical competency measure, a dynamic criminal knowledge base and visualisation to provide an investigator with an in-depth understanding of the case. The paper also describes how its implementation within a cloud-based infrastructure will also permit a more timely and cost effective solution.

Keywords
Digital Forensics, Computer Forensics, Artificial Intelligence, Cybercrime, Automation

INTRODUCTION

Whilst the rising use of technologies, such as the Internet, has brought the world closer, they have also provided a vast opportunity for criminal activities to be undertaken. An analysis of the trends within cybercrime have shown a consistent rise, with a study suggesting that they have increased 100% in the past 3 years alone. It is anticipated that this increase will continue and the world will certainly see a rise in cybercrime focussed upon the rising use of mobile devices and the increasing use of the Internet on such devices (Norton, 2012).

Some of the recent surveys and reports conducted by Norton (2013), McAfee (2013), RSA (2012), Ernst & Young (2011), and Ponemon Institute (2012) all indicate that cybercrime will certainly pose increasing challenges to digital forensics in the near future. Challenges such as:

- Threats due to Virtualization, Cloud Computing;
- Rising financial malware;
- Developing parallel black cyber economy using such tools;
- Fraud as a Service;
- Risk management investment;
- Rising in insider threats.

The RSA 2012 report states that cybercriminals are becoming more equipped with sophisticated technology by the day. Software packages such as Zeus are emerging as hugely popular tools in the Internet black market, which has advanced algorithms to break security and used in financial crimes and frauds. Symantec's Internet Security Threat Report (2013) provides a usful illustrated as to he nature and scale of the problem: a 42% increase in targeted attacks; 5,291 new vulnerabilities; 2.3 million bot-infected computers and a 58% increase in mobile malware families.

Therefore, the field of digital forensics is facing new challenges in the face of rising cybercrime, expanding use of the internet, rising volumes of data and information, and varied devices being used (Hunton, 2009). Unfortunately, under these circumstances, the time taken to undertake a case and the human-effort required is only increasing. This means that forensic examiners have to be given more effective tools that allow them to more rapidly identify relevant artefacts from the huge volume of noise that exists.

The paper proposes the utilisation of advanced automation techniques to develop an intelligent system that is able to identify, map and correlate artefacts within a case. The use of automation is already widely utilised within forensics for processing and extracting relevant information. For example, the use of hashing to identify known and notable files, or file signature analysis for the identification of data hiding. However, such approaches to date are rather simple. The correlation of artefacts and the interpretation of the evidence is the sole responsibility of the forensic examiner. Within information security more widely however, the use of artificial intelligence (AI) to analyse, correlate and interpret large volumes of data has been exhaustively applied (O' Leary, 2013). The paper presents an Automated Forensic Examiner (AFE) that is capable of utilising AI and criminal profiling to identify, extract and correlate suspect data.

Section 2 presents a literature review of current research in the area of automation for digital forensics. Section 3 presents the concepts of the criminal profiling and technical competency – a key feature for determining the depth of an investigation. Section 4 5 present the Automated Forensics Profiler (AFE) and the accompanying operational architecture (AFE) with a detailed explanation about its function. A discussion of the proposed system is given in Section 6 prior to the conclusions and future work.

LITERATURE REVIEW

As previously highlighted, the concept of utilising automation is already widely utilised in digital forensics. However, the level and depth to date in operational systems has been rather simple. Automation can be also utilised as a triage function, enabling investigators to understand whether the case image is worth investigating – however, again, the level of functionality here is based upon simple string or pattern matching processes. More recently however, a number of researchers have been undertaking studies to develop more advanced automation strategies.

One of the approaches of automation is the CBR or Case Based Reasoning (Amadot and Plaza, 1994). To state in simple terms, the case based reasoning concept tends to provide solutions to the problem based on its knowledge base, which is fed into it using previous investigations. The CBR approach heavily depends upon the facts of information stored in the knowledge base, which in turn are stored in the form of abstract information and not complete solutions. Whilst CBR makes an attempt to identify relevant artefacts, it is not capable of appreciating the relationship between them. It therefore still requires a human investigator to provide this correlation. Hence, this technique may not be suitable under all circumstances. There needs to be further research in the field where the knowledge base is developed in a systematic approach and that the tools are frequently checked to ensure that the output from the CBR system is the same or similar to that given by a human forensic expert.

Gladyshev and Enbacka (2007) provided an automated method for tracing such irregularities and inconsistencies where deliberate attempts have been made to tamper with the normal log files to hide trace artifacts. Proposed as the B-Method, the basic principle underlying this automation attempt is that although a user could alter information locally or remotely, it is not always possible to do this in a consistent manner. Since multiple data structures are involved in logging various activities, the perpetrator would most likely leave out some or other trace, and this inconsistency would be useful to pinpoint that some problem does exist regarding that data or log. Whilst certainly

very useful for incorporation within a wider system, the level of automation has been applied to a very specific forensic analysis.

FACE or framework for automatic evidence recovery and correlation was another good attempt by Case et al (2008) where the researchers developed a solid framework for the purposes of automation and also presented a tool called “ramparser” for automation in Linux based systems. Ramparser creates a memory dump and analyses it for relevant information (such as network connections and user activity). However, again, this automation effort is focused upon a specific analysis – which, while useful is not an approach that can be more widely applied. Getting relevant information about various running processes and applications is merely one part of the investigation.

There are other tools which can perform similar functions, but these lead to a fragmented picture of different sources of information, with hardly any apparent link with each other. This means that the investigators will still need to work hard to find the missing links in trying to create a bigger and more complete picture.

Whilst some efforts are being made to partly automate processes thus helping to save time and resources, approaches to date focus upon specific analyses and fail to incorporate more advanced AI-based approaches. Indeed, Casey & Friedberg (2006) believe that it is not easy to fully automate the entire digital forensic examination process largely due complexity and the current level of capability within machine learning. Therefore, they suggest automation can mainly be applied to routine tasks rather than tasks requiring intelligent reasoning like human investigators are capable of doing. Whilst there certainly is a question of how intelligent these AI approaches can be, there wide use within other areas of computer science and information security, suggest they would have a positive contribution. Clarke N. & Furnel S. (2006)**.**

Without this level of automation, the process of digital forensics would not stand a chance to survive the onslaught of the immense number of cybercrime incidents and the growing volumes of data they have to deal with. However, triage tools also have certain limitations, which need to be overcome, and this has to be achieved through the process of automation.

CRIMINAL PROFILING

The basic fundamental concepts of cyber profiling are based on the premise that common factors exist within cybercrimes and cyber criminals. For example, in child pornography cases would typically involve image-based evidence, while bribery cases would involve some level of communications-based evidence. Researchers have tried to build a system of detecting the perpetrators by taking note of some of the common factors within a crime scene, a criminal action, or through modelling the characteristics and motivations of the crime (Arthur et al, 2008). The process of identifying evidence normally consists of monotonous and laborious processes of scanning the entire data set of suspected material and an automated process would be best suited for such repetitive work by sorting, arranging and searching of items against some known parameters.

The concept of profiling existed long before cybercrime or cyber criminals were even heard of; however, the basic concepts of such profiling are not overly different from what the modern day profiling of cybercrimes and cyber criminals (Horsman et al, 2011). There have been various attempts to build frameworks to tackle cybercrimes and bring cybercriminals to justice based on the identification of common factors between them (Hunton, 2009), but to date, much of this research exists outside the domain of digital forensics in the area of criminal psychology. Little research has linked high-level criminal features to low-level computing-based objects.

The purpose of this research was to investigate from other domains such as criminal psychology what features exist that indicate themselves to be criminal and to develop a series of models that would assist in mapping and identifying evidence through the use of artificial intelligence-based systems. Artefacts would be correlated within the “intelligent system” to develop a holistic evidence

locator and collector. As illustrated in the Figure 1, the proposed approach utilises an iterative-based approach to identify evidence and then perform associative mapping to related events. It is anticipated that this approach would enable the system to create "evidence trails" linking together a series of related events, which would give rise to additional artefacts. In this manner, it will be possible to build up an understanding of actions a user undertakes. Whilst literature exists to demonstrate how crimes can relate to very simple computer objects (e.g. child pornography typically maps to image-based artefacts), the novelty in this work is the creation of relevant evidence trails and in the filtering and refining processes to reduce the effects of noise.

Figure 1: Automated Evidence Profiler

As illustrated in Figure 2, once initial artefacts have been identified through the simple crime-mapping to artefacts, the AFE automatically creates a series of chronology trials of the artefact – each chronology based upon a context within which it was used (i.e. within the file system, within email, or an attachment within a Skype call). Through mapping all activities prior to and after using the artefact, the system is searching for further artefacts that pertain to the case. The premise of the approach is based upon the concept that in order to use the artefact in the first instance, the suspect must be undertaking a series of actions that pertain to that activity. Therefore, it seems logical the suspect machine will have a series of criminal and normal evidence trials and the purpose of the AFP is to identify and extract the criminal ones. Moreover, correlations between the identified artefacts will be undertaken – those with high degrees of correlation will refer to artefacts that have a higher probability of being pertinent and thus are prioritised.

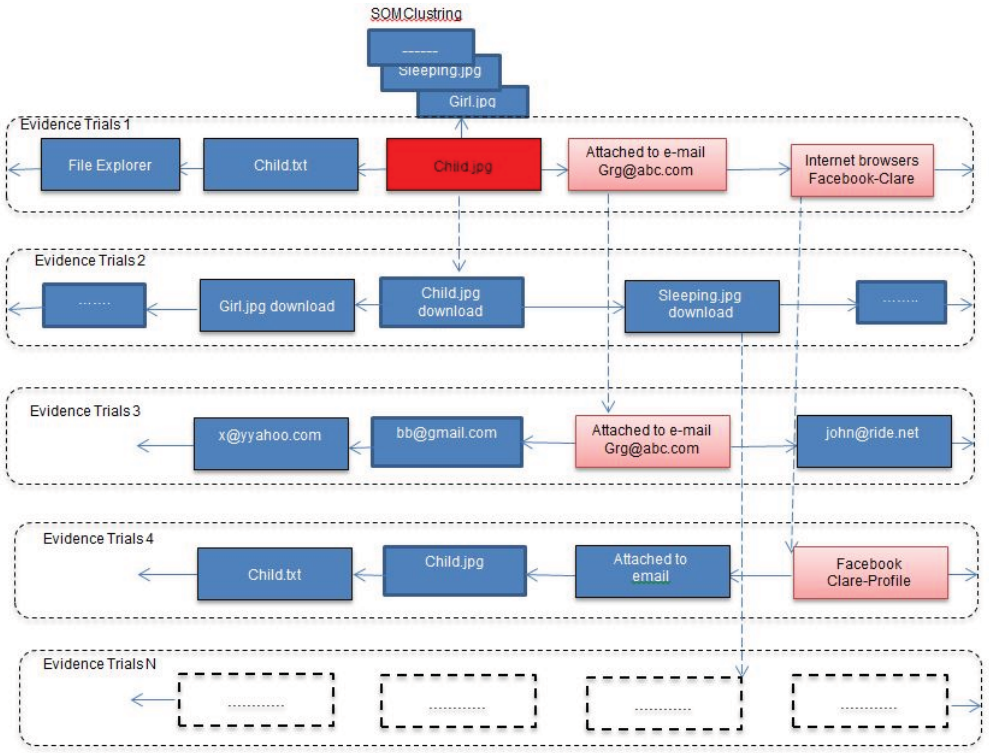

Figure 2: Example of Evidence Trials

Technical Competency

The time taken to examine a case (automated or otherwise) will be dependent upon the depth of analysis required – with systems belonging to suspects that have a limited knowledge of computing systems (and in particular data hiding) requiring a differing level of analysis to machines whose suspects have advanced technical competency to modify, hide and obfuscate their actions. The purpose of this process is to augment the criminal profiling approach through determining a measure of the technical competency of the suspect.

Criteria have been developed that can have an impact upon technical competency. For example, the presence of anti-forensic applications on a system would highlight a suspect with at the least sufficient knowledge of what such applications enable. Modifying or changing basic configuration options such as the sector size would also provide intelligence that the suspect has been modifying settings, possible to the advantage of hiding data. Table 1 provides an overview of the criteria; with an associated impact level indicating the degree to which or the weight that criterion has within the overall measure.

Table 1: Technical Competency Criteria

Criteria	**Impact**
OS Base Configuration (cluster and sector size, MFT core file manipulation)	High
Software development environments	Medium
Information security tools	High
Hacking/exploitation tools	High
Anti-Forensic Tools	High
Empty Recycle Bin	Low
Encryption	Medium

Criteria	Impact
Wiping software	High
Database software	Low
Deleting the log	High
Clearing browsing history	Low
Proxy servers	Medium
Steganography software	High

The technical competency would help insure that the desired level of analysis would be considered and that no potential evidence had been missed or ignored. On the other hand, if this measure indicated that the suspect was naïve or an ordinary user, more advanced analyses would not be invoked within the AFP and only evidence found during the normal analysis would be passed on for processing.

AUTOMATED FORENSIC EXAMINER

In order to realise the AFP and Technical Competency (TC), it is necessary to design an architecture that could support the aforementioned processes. As illustrated in Figure 3, the architecture comprises of a number of key processing stages: Forensic Pre-Processing, AFP, TC, Visualizer, Profiler Refiner and Report; and data storage elements.

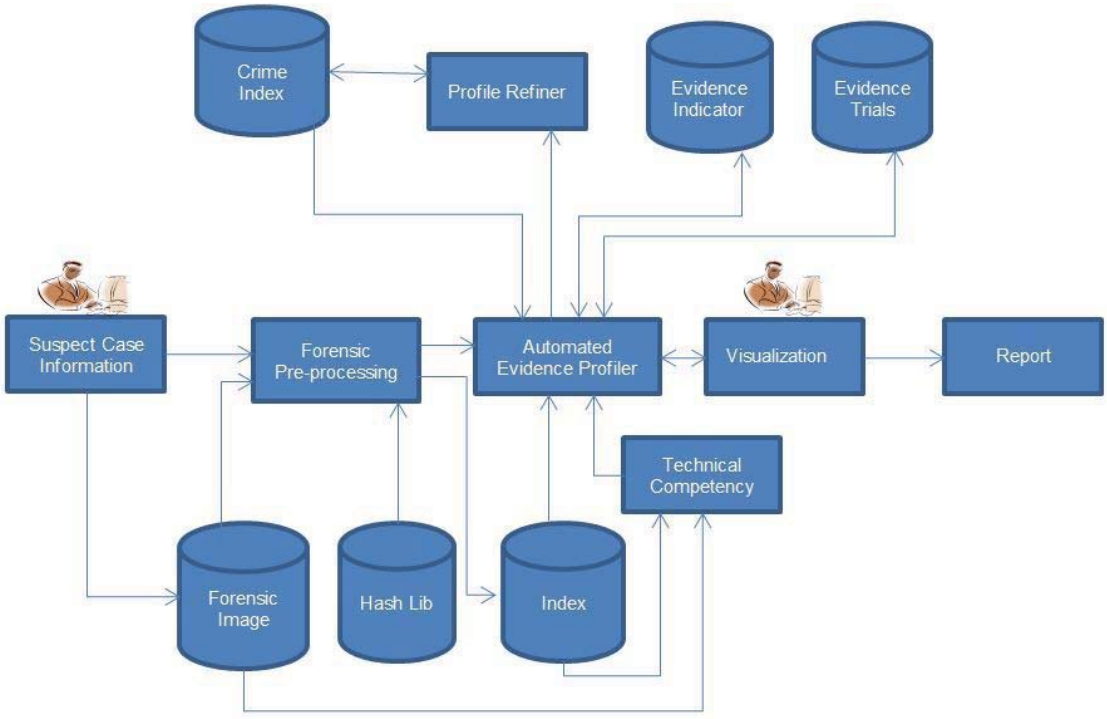

Figure 3: AFE Architecture

At the Suspect Case Information process, all the available suspect and case information would be fed into the system by the investigator. This is based on the assumption that the suspect is known and that the device used to carry out the attack has already been seized and an image acquired. The Forensic Pre-Processing stage will undertake a variety of standardised forensic process upon the image – including a hash analysis for known and notable files, files signature analysis, extraction of compound files, data and meta carving, keyword searching (based upon entered suspect information and pre-defined search criteria) and indexing. The primary role of this process is to reduce that

dataset and effectively sort the “wheat from chaff” in a manner that the relevant information gets separated from unnecessary information. Indexing permits parsing of the data so that it gets stored in a manner that makes information retrieval efficient later on. In the absence of such indexing, it would consume unnecessary time and computing power to search for any specific data items. Parsing tools and techniques have been used earlier in efforts to develop automated forensic tools by different researchers (Abbot et al, 2006; Case et al, 2008; Schatz et al, 2006).. In the case of the AFE, it would not be possible to apply “intelligent” parsers to the data prior to establishing a complete index.

Through indexing, the AFP is provided with an ordered and reduced dataset from which to perform its analysis. Prior to doing so however, the TC process is utilised to appreciate the type and level of analyses being undertaken. Through an analysis of the complete image (as standard programme files and data might be removed via the hashing process) TC will provide a list of advanced analyses that need to be undertaken depending upon the identified criteria. Notably, it will also provide an overall measure for technical competency in order to provide the investigator an appreciation of the case complexity.

The Automated Evidence Profiler is the core component of the Automated Forensic Examiner (AFE) and is the place where are the activity associated with the mapping of the artefacts to evidence trials occurs. The different types of data including but not limited to graphics, text, audio, timestamps, contacts, email communications, browser behaviour is mapped and updated to make a profile of the information within the case. Further, more advanced analyses will also be undertaken depending upon the outcome of the TC analysis. The Crime Index database contains the criminal profiling knowledge base. Whilst initially stored with well-accepted crime-artefact mapping information, this database will evolve over time to include patterns of behaviour from prior cases. Through the Profile Refiner, this permits the system to adapt to the changing cybercrime environment as new terminology and artefacts are created.

The Evidence Indicator database stores the extracted artefacts that the AFP process has identified; thus presenting a centralised collection of evidence pertaining to the case. The Evidence Trials database is utilised to store the metadata associated to the extracted artefacts.. The Visualizer represents the link between the AEP and the final report output. Recognising that the AFP process will inevitably identify false evidence trials and thus artefacts, this process exists to conveniently and useably present the evidence trials so that an investigator can discount or decrease/increase the priority of the trials. The Reporting process is the final output of the AFE that represents the analysis and the results of the entire investigation exercise.

DISCUSSION

Despite the fact that the past attempts of overcoming such a problem through triage and partial automation have enhanced the digital forensic domain, the need for a comprehensive automation system is vital to meet the future requirements of the domain. Criminal Profiling is one approaches to study the criminal characteristics and motivations which when used in the long term can provide the investigator with a rich database of useful information that can be used in future cases, thus reducing the time taken to prove or otherwise the case.

The proposed Automated Evidence Profiler (AEP) features an iterative-based approach to identify potential evidence and perform associative mappings to related events which enables the system to create evidence trails that is able to filter and refine the processes. The evidence trails are created to provide an atefact mapping through linking the related events together. For example, if the case was about child sexual abuse and a relevant image was found, the system would trigger an in depth search to find more similar images. Another example of this feature is that if the suspect had deleted some record, this would trigger trials surrounding the use of that artifact, with the intention of locating further artifacts (whether they be images or information pertaining to other offenders or

what the suspect used them for).

In order to undertake the analysis, Artificial Intelligence (AI) techniques such as the SOM (Self-Organizing Maps) will be utilised to better understand the data and the relationship between artefacts. Clustering has been used extensively to effectively organise large volumes of data by grouping related-events into smaller number groups (Kohonen, 1990). This mechanism provides the AFE a mechanism to effectively sort the events and provide information into the creation and correlation of Evidence Trials.

Digital forensic analysis is already a computational intensive task with pre-processing of large images taking many hours to complete. The introduction of further processing stages will only seek to extend that requirement. It has therefore been decided to implement the AFE within a cloud-based Infrastructure as a Service (IaaS) platform in order to take advantage of the scalable and dynamic processing environment. This centralized service will provide more timely analysis, be in a position to benefit from case history and thus updates to the criminal profile knowledge base. A web-based front-end to the visualization and reporting processes will also ensure access to the results can be independent of specialist forensic software and platforms – further reducing the cost.

CONCLUSION & FUTURE WORK

The proposed approach in this paper aims to address a significantly growing gap between the number and size of cases that require forensic examining and the time taken for investigators to process each case by enhancing the analysis process through introducing advanced levels of automation. The proposed solution consists of a number of key processes that permit advanced analysis (Technical Competency and Automated Forensic Profiler), adaptability through the Profile Refiner and a feedback mechanism through the Visualizer.

Incorporating this within a cloud solution, that can adapt dynamically to the resources required, including the parallel analysis of multiple cases, provides a solution that at least will enable the identification of images that require further examination by a human-based investigator but also offers up the opportunity to begin in certain situations to remove the need for an investigator. Freeing up valuable expertise to investigate more complex cases.

The AFE is currently under implementation and future work will focus upon developing a scientific validation for the approach. Whilst empirical proof will be difficult to establish due to the nature and complexity of the cases, a real-world evaluation will be performed through access to a historical database of previous cases provided to the authors. A technical evaluation of the cloud-based system will also be undertaken to understand the time and cost benefits of utilising such a platform for forensic processing.

REFERENCES

Aamodt, A. Plaza, E. (1994); Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications. IOS Press, Vol. 7: 1, pp. 39-59.

Arthur, K.K., Olivier, M.S., Venter, H.S. and Eloff, H.P. (2008) Considerations Towards a Cyber Crime Profiling System Information and Computer Security Architectures (ICSA) Research Group DOI 10.1109/ARES.2008.107

Carrier, B. &Spafford, E.H. (2004) Getting Physical with the Digital Investigation Process.International Journal of Digital Evidence, Fall 2003, Volume2, Issue2

Carrier, B. (2006) Defining digital forensic examination and analysis tools using abstraction layers Vol. 1 (4) [online] Available at http://www.digital-evidence.org/papers/dfrws_define.pdf (Last accessed 01 Oct 2013)

Case, A, Christina, A, Marziale, L, Richard, G, G, Roussev, V 2008, 'FACE: Automated digital evidence discovery and correlation' ,Digital Investigation, 5, S65-S75, ELSEVIER, Science Direct, doi:10.1016/j.diin.2008.05.008

Casey, E. and Friedberg, S. (2006) Moving forward in a changing landscape Digital Investigation 3, 1-2

Clarck, N. and Furnell S. (2006) Authentication Mobile Phone Users Using keystroke Analysis. International Journal of Information Security. Volume 6, No.1, pp1-14

Ernst & Young (2012) Cybercrime diagnostic: Pro-actively combating high-impact cyber threats. Available at http://www.de.ey.com/Publication/vwLUAssets/Cybercrime-diagnostic/$FILE/1367006_FIDS_Cyber_diagnostic_Flyer_UK_4_DRAFT.pdf

Gladyshev, P. and Enbacka, A. (2007) Rigorous Development of Automated Inconsistency Checks for Digital Evidence Using the B Method. International Journal of Digital Evidence 6 (2).

Horsman, G. Liang, C. and Vickers, P. (2011) A Case Based Reasoning System for Automated Forensic Examinations. In: PGNET 2011 The 12th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting 27-28 June Liverpool.

Hunton, P. (2009) The growing phenomenon of crime and the internet: A cybercrime execution and analysis model. Computer Law and Security Review, 25, s28-s35.

ICSPA (2012) The Impact of Cybercrime on Canada. [Online] Available at: https://www.icspa.org/uploads/media/ICSPA_Canada_Cyber_Crime_Study__ROW__-_Media_Release__Final__01.pdf (Accessed: 30 September 2013)

Kohonen, T. (1990) The self-organizing map. IEEE Proceeding (Volume:78, issue:9)

Lim, S. Savoldi, A. Lee, C. and Lee, S. (2012) On-the-spot digital investigation by means of LDFS: Live Data Forensic System Mathematical and Computer Modelling Volume 55 223 - 240

McAfee (2013) The economic impact of cybercrimes and cyber espionage. Available at http://www.mcafee.com/uk/resources/reports/rp-economic-impact-cybercrime.pdf

Norton (2013) CyberCrime report 2012. Retrieved from: http://now-static.norton.com/now/en/pu/images/Promotions/2013/PDFs/NCR%20-%20%20Mobile%20-%20Europe%20FINAL%20FINAL.pdf

O'Leary, D. E. "Artificial Intelligence and Big Data," IEEE Intelligent Systems, vol. 28, no. 2, pp. 96-99, March-April 2013, doi:10.1109/MIS.2013.39

Ponemon Institute (2012) The impact of cybercrime on business: Studies of IT practitioners in the United States, United Kingdom, Germany, Hong Kong and brazil. Retrieved from http://www.ponemon.org/local/upload/file/Impact_of_Cybercrime_on_Business_FINAL.pdf

Ponemon Institute (2012) Cost of cybercrime study: United States. Retrieved from http://www.ponemon.org/local/upload/file/2012_US_Cost_of_Cyber_Crime_Study_FINAL6%20.pdf

RSA (2012) Cybercrime trends report: The current state of cybercrime and what to expect in 2012. Retrieved from http://www.rsa.com/products/consumer/whitepapers/11634_CYBRC12_WP_0112.pdf