2016

Automated Digital Forensics and Computer Crime Profiling

Al Fahdi, Mahmood

Automated Digital Forensics and Computer Crimes Profiling

By

Mahmood Al Fahdi

A thesis submitted to the Plymouth University in partial fulfilment for the degree of

Doctor of Philosophy

December 2015

COPYRIGHT STATEMENT

Abstract

Automated Digital Forensics & Computer crime Profiling

Mahmood Al Fahdi

Over the past two decades, technology has developed tremendously, at an almost exponential rate. While this development has served the nation in numerous different positive ways, negatives have also emerged. One such negative is that of computer crime. This criminality has even grown so fast as to leave current digital forensic tools lagging behind in terms of development, and capabilities to manage such increasing and sophisticated types of crime. In essence the time taken to analyse a case is huge and increasing, and cases are not fully or properly investigated. This results in an ever-increasing number of pending and unsolved cases pertaining to computer crime.

Digital forensics has become an essential tool in the fight against computer crime, providing both procedures and tools for the acquisition, examination and analysis of digital evidence. However, the use of technology is expanding at an ever-increasing rate, with the number of devices a single user might engage with increasing from a single device to 3 or more, the data capacity of those devices reaching far into the Terabytes, and the nature of the underlying technology evolving (for example, the use of cloud services). This results in an incredible challenge for forensic examiners to process and analyse cases in an efficient and effective manner.

This thesis focuses upon the examination and analysis phases of the investigative process and considers whether automation of the process is possible. The investigation begins with researching the current state of the art, and illustrates a wide range of challenges that are facing the digital forensics investigators when analysing a case. Supported by a survey of forensic researchers and practitioners, key challenges were identified and prioritised. It was found that 95% of participants believed that the number of forensic investigations would increase in the coming times, with 75% of participants believing that the time consumed in such cases would increase. With regards to the digital forensic sophistication, 95% of the participants expected a rise in the complexity level and sophistication of digital forensics. To this end, an automated intelligent system that could be used to reduce the investigator's time and cognitive load was found to be a promising solution.

A series of experiments are devised around the use of Self-Organising Maps (SOMs) – a technique well known for unsupervised clustering of objects. The analysis is performed on a range of file system and application-level objects (e.g. email, internet activity) across four forensic cases. Experiment evaluations revealed SOMs are able to successfully cluster forensic artefacts from the remaining files. Having established SOMs are capable of clustering wanted artefacts from the case, a novel algorithm referred to as the Automated Evidence Profiler (AEP), is proposed to encapsulate the process and provide further refinement of the artefact identification process. The algorithm led to achieving identification rates in examined cases of 100% in two cases and 94% in a third.

A novel architecture is proposed to support the algorithm in an operational capacity – considering standard forensic techniques such as hashing for known files, file signature analysis, application-level analysis. This provides a mechanism that is capable of utilising the A E P with several other components that are able to filter, prioritise and visualise artefacts of interest to investigator. The approach, known as Automated Forensic Examiner (AFE), is capable of identifying potential evidence in a more efficient and effective manner. The approach was evaluated by a number of experts in the field, and it was unanimously agreed that the chosen research problem was one with great validity. Further to this, the experts all showed support for the Automated Forensic Examiner based on the results of cases analysed.

Contents

List of Figures

List of Tables

Acknowledgement

The successful completion of this doctoral thesis was possible with the support of many people. I would like to express my sincere gratitude to all of them. First of all, I would like to thank my Director of Studies Professor Nathan Clarke who has been always supportive to me during my PhD study; without his instructional guidance and endless support, this thesis would not have been possible. Also, I would like to thank Professor Steven Furnell for being my second supervisor. His rich experience and knowledge in the field played a major factor towards the success of my research project. My thanks also go to Dr Fudong Li for his support and motivations during my PhD journey.

I would also like to thank my employer who has generously funded me along my high education journey. Many thanks as well to the head of e-crimes investigation in the Public Prosecution in Oman, Mr. Saeed Al Muqbali and his assistant Mr. Abdullah Al Kharousi for their support during the research experiment phase.

Special thank dedicated to His Excellency, Sayyid Saud bin Hilal Al Busaidi, Minister of State and Governor of Muscat for the extraordinary support and priceless help and advices to overcome many difficulties during my PhD research.

I'm very much indebted to my family, my parents, wife, children, brothers and sisters who supported me in every day during my PhD research journey. With the great support from those I mentioned, the completion of this work was possible.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Sub-Committee.

Work submitted for this research degree at the Plymouth University has not formed part of any other degree either at Plymouth University or at another establishment.

This study was financed with the aid of a scholarship from the Government of the Sultanate of Oman.

Relevant seminars and conferences were attended at which work was often presented and several papers prepared for publication.

Word count of main body of thesis: 58,420 words

Signed ……...

Date ……...

1 Introduction

This chapter starts by introducing the scope of the research and the problem that is to be examined during the course of this document. Then, it highlights the aim and the objectives of the research, followed by a summary of each chapter and the key points contained within it.

1.1 Introduction to the Research Domain

Advancements in modern technology have a substantial impact on our daily existence. Communication networks are highly developed, enabling faster and easier exchange of information. Statistics with respect to penetration of technology in various parts of the world indicate a significant rise in the number of devices and the use of the Internet. The gap between the developed and the developing countries, with respect to the technology, is continuously shrinking and has become minimal as far as utilization of the technology is concerned (ITU, 2012a). The role of the technology is also evident from several other aspects. Indeed since the first domain name was registered in 1985, the Internet has seen an exponential growth. Currently, the Internet economy generates nearly 2 to 3 trillion dollars annually but nearly one-fifth of that is subsequently attributed to computer crimes[1] (McAfee, 2014). The economic impact of the Internet is growing bigger. As demonstrated in Figure 1.1, the International Telecommunication Unit (ITU) statistics suggest that the number of Internet users increased significantly from 400 million in 2000 to 3.2 billion in 2015, an 8-fold increase.

Figure has been removed due to Copyright restrictions.

Figure 1.1: ITU statistics for the Internet users (2000-2015)

Whilst the rising use of the technology such as the Internet has brought the world closer together, and has provided the opportunity for individuals and businesses to grow, negatively it has also provided a platform where new forms of criminal activities are able to take place (Salifu, 2008). With respect to crimes, computer crime statistics and studies including (but not limited to) Rotich et al (2014), ONS (2014), CFS (2012) and NCA (2014) indicated that

[1] Computer crime is defined as "any illegal act involving a computer, its system or its application and the Internet" (EC- Council-Press, 2012)

incidents have increased rapidly and the impact on individuals, companies and nations has correspondingly grown significantly. The list of such cyber incidents is countless and ranging from targeted attacks on user login credentials via phishing for financial purposes, to obtaining access to and publishing private photographs of celebrities (Grabosky, 2004; Harnish, 2015). Extensive surveys conducted by governments such as the National Audit Office (NAO, 2013) and companies such as PricewaterhouseCoopers (PWC, 2014) all pointed towards the seriousness of the situation, indicating the necessity for effective measures being required to mitigate cyber risks.

Figure 1.2 shows the Internet Crime Complaint Center's (IC3) statistics on computer crimes on a monthly basis during 2014 (IC3, 2014).

Figure (Text/Chart/Diagram/image etc.) has been removed due to Copyright restrictions.

Figure 1.2: Stats of Rising Computer crime (IC3, 2014)

Furthermore, the Ponemon Institute conducted a survey pertaining to computer crime cost in seven countries in 2013 and 2014 (as demonstrated in Figure 1.3). In 2013, a total of 235 countries participated in the survey; and 257 countries took part in 2014. These numbers clearly reflect the magnitude of computer crime and its impact (Ponemon, 2014).

Figure (Text/Chart/Diagram/image etc.) has been removed due to Copyright restrictions.

Figure 1.3: Total cost of computer crimes in seven countries (Ponemon, 2014)

There are several definitions for digital forensics science. However, the most common used is the definition that was given by the Digital Forensic Research Workshop (DFRWS) as *"The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations*" (DFRWS, 2001).

Computer crime has undergone a transformation, evolving from fragmented acts into a highly organized, sophisticated and professional activity (Chouhan, 2014). It is fast expanding to industrial proportions where professional networks undertake illegal operations (Knapp et al, 2006; Tropina, 2012). The debate regarding cyberspace as a ground for criminal activities covers two main aspects – use of cyberspace by traditional organized groups and use of cyberspace for new forms of organized crimes (Tropina, 2012). In fact, the scale of the

problem has reached a point where parallel underground economies exist in cyberspace, operate at various organized levels and deal with different types of crimes (e.g. credit card fraud and money laundering) (Symantec, 2015). Baar et al. (2014) states that the development of cyber forensics is an important factor in curbing the rise of computer crime.

Broadly speaking, the challenges faced by digital forensic examiners have been classified under three headings (Marcella and Greenfield, 2002): namely technical, legal and resource challenges; each of these will be discussed in detail in the later sections of this thesis. To get a brief overview of the types of such difficulties, they could range from jurisdiction issues (Wilson, 2008; Cohen, 2009) and continuous development of new tools and techniques to keep pace with cyber criminals (Mohay, 2005), to non-uniform availability of resources to fight such crimes in developing regions of the world (Gercke, 2011). It is important that proper tools and techniques are developed to overcome the ongoing challenges faced when dealing with such crimes.

The preceding literature points towards the need to devise innovative tools and techniques which can relieve the rising burden from digital forensic examiners. This thesis focuses upon the application of artificial intelligence to develop the next generation of computer forensic capability.

1.2 Aim and Objectives

The main aim of the proposed approach is to reduce the time taken to investigate a computer assisted crime. In order to achieve this, it was imperative to research the current state of the art and address the identified gap with an appropriate solution in a scientific manner. Thus, the following objectives were setup as the research plan:

1. To establish the current scale of computer crime (with a special focus upon computer crimes) and the degree to which digital forensics is relevant to support investigations, by researching the history of computer crimes and drawing a correlation between technological development and the rising rate of computer assisted crimes.

2. To research existing digital forensic methodologies and computer forensics tools that are used to extract and examine computer evidence, and then to determine the requirements for the next generation tools.

3. To investigate the challenges faced by digital forensic examiners, and to establish a link between researcher's views in this regard with the challenges faced in reality.

4. To design a novel approach to the identified gap in a scientific manner that would contribute to solving the given problem.

5. To test the designed algorithm with data collected from a public domain and real world cases.

6. To evaluate the functionality and the accuracy of the designed approach by seeking opinions and feedback from experts within the digital forensic field.

1.3 Thesis Overview

The initial chapter starts with an introduction to the domain and illustrates the rapid developments in the technology. It discusses the rise of computer crime and how it has become a menace for most countries and evolved over time from isolated incidents, to being targeted by highly sophisticated and organised crime syndicates. Finally the chapter explains why the development of an automated forensic investigation system has become a necessity in current times and how such a system could help digital forensic investigators and agencies across the world.

Chapter two discusses the history of computer crimes and explores associated trends such as the rising use of the Internet and networked technologies, and the increasing rate of computer crimes. The impact of computer crimes on society at different levels of abstraction such as government, corporations and individuals will be assessed based upon statistics related to the technology trends and its impact on users.

Chapter three examines the digital forensic methodologies and discusses their developments, specifically towards the beginning of the 21st century. A comparison between the different methodologies and frameworks that are used in the digital forensic domain is discussed in breadth and depth. Based on the detailed analysis of various methodologies, the role that

automation plays within investigation is clearly identified: automation is a necessity for digital forensic examiners rather than an optional extra.

Chapter four reviews a number of popular first generation digital forensic tools in terms of their features and capabilities. The selected tools are termed as first generation because most of them are only capable of performing evidence recovery and are not suitable to undertake actual digital forensic examination. A comprehensive identification of the limitations of first generation tools leads to the assimilation of the desired characteristics of second-generation tools. The concept of triage is also discussed which deals with sorting evidence according to case background.

The fifth chapter focuses on the existing challenges faced by digital forensic examiners, including technical, legal and resource challenges. The chapter also lists several challenges that forensic investigators will most likely encounter in the future, and the importance of tackling those challenges. The chapter then outlines a survey which investigates stakeholders' views on various digital forensic challenges. The analysis of the results also highlights the trends across a range of parameters relating to digital forensics.

Chapter six initially describes the process of Self Organizing Maps (SOMs) and their use in the field of digital forensics. Then, a series of experiments are devised to explore the feasibility of SOMs upon analysing artefacts within a forensic case. The experimental results are analysed and discussed in detail to show the impact of various factors on the output.

Chapter seven presents the *Automated Evidence Profiler* (AEP) – an algorithm that combines the use of SOMs with a timeline analysis process to refine and optimise the identification of relevant artefacts. The concept and process of timeline analysis and evidence trails are also explained in detail. The chapter then describes a set of experiments that are used to investigate the effectiveness of the AEP algorithm. All the experimental results are then presented and discussed.

Chapter eight presents a novel *Automated Forensic Examiner* (AFE) architecture, which is incorporated with the AEP, to provide a firm foundation for an operational system. The functionality of various components is discussed as well as their contribution to the overall

architecture. The chapter also explains the concepts of computer crime (a branch of computer crime) profiling and the estimation of the technical competency of the suspect.

Chapter nine describes the evaluation process of the research. Within the evaluation, the theoretical concepts and the technological output (i.e. AFE) are analysed by six independent experts. The rationale behind the evaluation is also explained followed by the participant selection process. Details of the evaluation results are also presented and thoroughly discussed.

Finally, Chapter ten presents the main conclusions of the research along with the achievements and the limitations of the project. The chapter also contains a summary of potential future research directions.

2 Computer crimes

This chapter discusses the term computer crime in detail, including its impact and historical trends. Also, the link between the growth of computer crime and ever changing technologies will be explored; in addition the range of targets that may be affected, from individuals and businesses, to government organisations, will be described. Digital forensics is the primary technique by which a computer related crime can be examined and analysed. In order to appreciate the impact of digital forensics, a review and analysis of the current computer crime environment was required. This chapter details the history of computer crime and its connection to the technology revolution, whilst highlighting the impact of the computer crime on society.

2.1 Introduction

Before examining the history of computer crime it is important to understand what exactly is meant by the term computer crime. In the most basic language, computer crime can be explained as a criminal activity which is perpetrated with the help of any electronic devices such as office computers, mobiles, laptops etc. The range and nature of the activities that can be classified under the heading of computer crime is vast and includes but not limited to identity theft, cyber bullying, financial scams, unauthorized access to confidential data, phishing, spamming, attacks on vulnerable or important computer systems, child abuse, and money laundering (Richet, 2013). Computer systems can either be the direct target of an attack or used as a resource to perform harmful and illegal activities (Parker, 2007).

The nature of loss and the consequences resulting from computer crimes vary from case to case. Whereas an individual or an organization can suffer huge financial losses due to crimes like identify theft or stealing of confidential data, there can also be non-tangible losses and harm such as intense mental trauma or even the compromise of national security (Parker, 2007; Kramer and Starr, 2009). With the rise of sophistication of computer crimes over time, the associated consequences have greatly increased in range and number. Computer crimes can also involve crimes of a heinous level like terrorist activities and attacks upon important security related systems being used by government bodies (Kramer and Starr, 2009).

The alarming rising rate of computer crimes is certainly a growing issue. Over time, the technology of networked devices are dominating all aspects of life and business related activities: from the use of Internet banking to online shopping, the number and value of transactions that are conducted over cyberspace have increased enormously. Combine this with the fact that many people are not fully aware of the dangers or implications of the improper use of such technologies and this makes them vulnerable, and presents them as potential subjects and subsequent victims of computer crime.

Since the individual cannot be separated from society, the rise of computer crime affects all levels of abstraction in society: from national governments, corporations, and organizations, right down to the individual. The implications of this impact can be better understood if consideration is given to how technology has touched each and every aspect of individuals and business entities. Nowadays it is possible to cripple the economy of an entire nation by targeting the networked infrastructure and the services that are available online (Betz, 2012).

Given the above scenario, it is clear how serious the situation is in terms of computer crimes and their possible implications. The following sections begin by exploring the history of computer crime from several decades ago when networking technologies were basic when compared with the current situation, and will trace the growth in level, sophistication and volume of computer crimes from that point forward.

2.2 Historical Trends of Computer crimes

During the earliest phases of computer related crime starting in the 1960s, the main targets of attack were the actual physical systems, and this remained the prominent case for approximately two decades until the emergence of networking. Starting from the isolated physical attacks on computer systems in these early days, crimes slowly began to grow in sophistication, from credit card frauds, spam, viruses, and Trojans, to the organised hacking of networks in the current era (Kabay, 2008).

Indeed, a hacker has been referred as the "archetypal 21st century criminal" (Burden et al, 2003) insofar as he or she seeks to gain unauthorised access to computer systems, with motives ranging from fun to serious misuse (e.g. stealing confidential or financial

information). As far as the law stands, the crime is the same irrespective of the motive behind the hacking activities.

As highlighted by Grabosky et al (2001), normally in any type of criminology, the basic principle is that crime follows opportunity. Since the growth of the Internet gave a platform for global interaction amongst people and computer systems, at an unprecedented level this also gave rise to an equal opportunity for criminals to utilise this opportunity for conducting illegal activities.

A survey conducted by Deloitte (2010) gave a clear indication that the sophistication level and the technologies used in computer crimes are rising at a rapid pace, and the hitherto universally accepted simple measures like antivirus etc. are fast becoming obsolete, especially for technically advanced hackers. They use different techniques such as encryption to avoid detection, and are becoming increasingly difficult to detect and recognise.

Moving from the historical perspective to relatively recent developments, one of the important developments that have taken place is the rising use of Cloud Computing with its associated vulnerabilities. Whereas the increase in the use of the Cloud Computing technologies is certainly helpful for individuals, businesses and organisations, it also raises risks and poses more challenges to law enforcement agencies (Grispos, 2011). With the prediction that nearly 40% of small and medium business organisations are expected to put their data in the Cloud in the coming years (Kazarian and Hanlon, 2011), this shows the level of risk to which this data could be exposed if proper security measures are not put in place.

Serious issues such as the rising use of the Internet for terrorist related activities have been observed in the last decade. For instance, the terrorist organization “Al Qaeda” was able to communicate and pass their plans to execute the September, 11 attack using secure mechanisms such as encryption and steganography. Such terrorist activists and groups have become more sophisticated and use encryption technologies to bypass security measures, even though most Internet Service Providers are careful about blocking websites related to terrorist organisations and access to information (UNODC, 2012).

While it would be a major task to list all the computer crime events that occurred during the past few decades, a description and taxonomy of some of the major events that happened would provide a better indication of the seriousness of the problem and its growth.

Starting at the beginning of the 1960s, the initial crimes for the first decade were more of a disruption rather than ‘computer crimes' in the truest sense of the word; rather, they were computer crimes. Later, the focus shifted more to the use of computers, or rather computer mediated crimes; and one such common type of crime was data diddling, which referred to unauthorised data alteration. One classic example of such a crime was the equity funding fraud, which occurred between 1963-1974, where an American equity company deliberately altered their computer data for nearly a decade to keep their stock prices rising and earned huge profits by selling fake policies (Kabay, 2008).

Keeping the growing computer crime menace in mind, the International Cyber Security Protection Alliance organised an extensive study to estimate the impact of computer crimes in Canada. The study commenced in November 2012, and culminated in the spring of 2013 (ICSPA, 2013). The organisation urged all companies to participate in their study, and the overall goal of the study was to judge the nature of cyber threats and their impacts upon Canadian businesses. Amongst the 520 businesses surveyed, a total of 5,866 computer crime incidents were reported (ICSPA, 2014). In addition, there were 8 mega breaches in 2013 as compared to only one in 2012. Mega breaches is a term used for data breach incidents that result in personal details of at least 10 million identities being exposed (Symantec, 2014).

Computer crimes may be triggered indirectly by factors such as unemployment amongst computer literate personnel who are unable to gain a respectable job or make a good living. The rising number of call centres in certain developing economies such as India, has also been blamed as one of the reasons for computer crimes because the chances of data theft in such regions are more than other places. In fact, such factors have given rise to a breed of career criminals who earn full or part of their income from these activities (Saini et al, 2012).

Computer crime has been growing not only in volume but also in the level of sophistication and seriousness, and there is a need to prevent or minimise such activities which are growing at alarming proportions in cyberspace. A survey conducted by PricewaterhouseCoopers (PWC, 2014) on rising risks and reduced readiness with respect to computer crime in the

United States, highlights that the sophistication of technologies being used in computer crimes is rising at a rapid pace. Simple measures like antivirus are fast becoming obsolete as cyber criminals are using advanced techniques like encryption to make detection more difficult.

Cloud Computing, a relatively recent development is a flexible and cost effective platform for providing solutions to businesses and individuals in the form of online storage, business applications, customised software and a realistic network environment (Chou, 2013). However, it comes with its share of vulnerabilities in the form of insecure interfaces and APIs, data loss and leakage, and hardware failure (Chou, 2013) - posing greater challenges for law enforcement agencies when an incident occurs.

2.3 Link between Technology Revolution and Computer crime

The widespread use of computer systems and information technology has revolutionised the way in which individuals and businesses interact and conduct business. Whilst such technology has added significant advantages for individuals, businesses and governmental sectors alike, it also created a larger target for potential victims who may not be fully aware of the threats of computer crime.

The level of sophistication for computer crimes has generally kept pace with the growth in technology. Early computer related crimes were mostly simple physical sabotage acts as described by Kabay (2008), where damage in the 1960s was mostly limited to the hardware or software of a local computer.

As technology grew, the number of people using an Internet enabled device has grown exponentially. In addition, the number of users subscribing to Internet services has also risen significantly. Indeed, according to a study by ITU (2012) it shows the annual growth rate of global Internet usage during the past decade, pertaining to the developing and the developed world:

- Cellular subscriptions tripled from 2005 to 2010, reaching nearly 6 billion.
- The growth percentage of developing and the developed world are nearing almost equal proportions.

- Broadband growth has been phenomenal across the globe, and this in turn has led to easy and cheap availability of the Internet to the masses.

Year	Growth Percentage of Developing World fixed line/broadband usage (per 100 people)	Growth percentage of Developed World fixed line/broadband usage (per 100 people)
2001	-/-	--
2002	100/100	100/34.4
2003	100/25	50/2.6
2004	20/60	33.3/9.7
2005	13.6/12.5	37.5/11.1
2006	20/11.1	18.2/4
2007	16.6/10	23/15.3
2008	14.3/18.1	12.5/3.3
2009	5/38.4	16.6/4.8
2010	7.2/11.1	9.5/6.1
2011	11.1/30	13/5.7

Table 2.1: The annual growth rates for Fixed Line Usage/B (ITU, 2012)

A study conducted by McAfee (2012), who gathered the insights of several respected experts in the field of computer security and digital forensics, highlighting the fact that computer crimes have risen gradually over time and will continue to do so. The rise of computer crimes will pose an increasing threat due to the higher sophistication and increasing organisational capabilities of cybercriminals. Rotich et al (2014) suggest that in the future, traditional street crimes might be history but computer assisted crimes will continue to rise due to the ease of perpetration and lesser risk of being caught or prosecuted.

According to a report from the Insurance Information Institute, the total number of data breaches and number of records exposed were the highest in the year 2013 (as illustrated in Figure 2.1) (Insurance Information Institute, 2014). The numbers of disclosed data breaches in 2011, 2012 and 2013 were 419, 449 and 614 respectively. Also, data breaches in 2013 exposed close to 92 million records. It is envisaged that the actual number of breaches is in fact much higher because a large number of attacks go unreported.

Figure (Text/Chart/Diagram/image etc.) has been removed due to Copyright restrictions.

Figure 2.1: Disclosed data breaches in 2013 (Insurance Information Institute, 2014)

The main reasons for this are the general lack of understanding towards cyber security, the use of very weak passwords and the use of services such as social networking with such a lax approach, which makes them lucrative hunting grounds for the cyber criminals. This can be rectified to a certain extent by raising general awareness about the loopholes which can be used to perpetrate computer crimes through such popular platforms (Verma & Sharma, 2014).

A similar study performed by RSA (2012) revealed that due to increasing sophistication and the rising availability of financial malware and blackhat tools in the underground cyberspace, Fraud as a Service (FaaS) has become a rising concept. A marked increase in the quality, quantity and complexity of computer crimes has been revealed in a survey carried out by PwC (2014). Such a fact concludes that cyber criminals target both private industry and critical infrastructure.

Therefore, it can be seen that from the humble beginnings of the 1980s with the rising use of computer technology in domestic and business sectors (Downland et al, 1999), to the modern Cloud Computing environments (Birk, 2011), the role of technology in computer crimes is increasing, and this technology has a positive correlation to various crimes (both new and traditional). It also follows that with the rising volumes of computer related crimes, there is a need for the development of new approaches to curb this growing menace (Sridhar et al, 2011).

2.4 Computer crime Impact

Computer crime is a menace which has been rising steadily, and it is having a major effect across all sectors and trends. It has been shown that computer crime rates have nearly doubled in the past three years (Computer Fraud and Security, 2012). Since the Internet is becoming an important part of all aspects of human society and its constituents such as the individuals, the businesses and even governments, the impact of such a rise has been felt across all these sectors. Table 2.2 shows the different recent surveys that point towards this growing menace of computer crimes.

Source	Participants	Theme/s	Observations	Misc.
Computer crime Report (Norton, 2012)	13,018 people from 24 nations	Scale, Price, Security IQ, Social Networking, Passwords	18 victims/Sec; $110 Billion loss; Cyber security; IQ is lacking; Passwords weak; Social network attacks arising	Russia, China and S. Africa top the list of victims
Prospective Analysis on Trends in Computer crime (McAfee, 2012)	Collection and analysis of 22 studies related to trends and projections from 2011 to 2020	Delphi method to find experts and get info via questionnaires related to different aspects of computer crime	In general, all experts agreed that in coming times, computer crime will be a major role in conventional crimes	Emerging threats will affect strategic data and also new inventions in IT would be misused for computer crimes
Cost of Cyber Crime (Ponemon, 2012)	2,618 IT professionals from business sector	Impact on businesses in US, UK, HK, DE and BZ	Germany has a higher computer crime loss Brazil has the lowest	Caveats associated with general surveys could be a limitation
Computer crime Trends Report (RSA, 2012)	Study by RSA Antifraud command centre	Trends in 2012 and beyond	Financial malware; Blackhat tools; Rising security threats	FaaS/Fraud as a Service concept in underground cyberspace
Information Security Breaches Survey (PWC, 2012)	447 organizations covering 10 sectors	To study corporate security breach incidents	93% large organizations had security breach; GBP110-250K cost per breach to large companies; GBP 15-30K for small companies	IT budgets need to focus more on security; Controls not keeping pace with innovation in technology
Computer Crime and Security Survey (CSI, 2011)	351 security practitioners from 9 industrial sectors	Open ended questions to gauge knowledge about cyber threat awareness	Malware infection most common attack Reluctance to give financial loss figures	Half of organizations don't use Cloud Computing as yet
Internet Security Threat Report (Symantec, 2011)	Based on data collected from Symantec Global Intelligence Network	Provide a comprehensive database of internet threats, attacks, malware identity theft.	5.5 billion attacks stopped in 2011; 1.1 million identities exposed per breach; 81% year on increase from the previous year	50% bigger and 18% small businesses attacked Mobile phones under growing attack

2013 Internet Crime Report (IC3)	FBI & National White Collar Crime Center	Internet related complaints across a variety of demographics	Total complaints: 262813; Total losses reported: $781,841,611	Top Five Countries ranked by the Total Number of Complaints Received by IC3 in 2013 – US, Canada, UK, India, Australia
Prospective Analysis on Trends in Computer crime (McAfee, 2014)	Collection and analysis of 22 studies related to trends and projections from 2011-2020	Delphi method to find experts and get info via questionnaires related to different aspects of computer crime	In general, all experts agreed that in coming times, computer crime will be a major role in conventional crimes	Emerging threats will affect strategic data and also new inventions in IT would be misused for computer crimes
Global Information Security Survey (Ernst & Young, 2014)	1,700 people across 52 nations Annual survey by invitation	External threats Rising use of virtual environments Cloud Computing Security policies	External threats arising Current security not adequate Cloud Computing Increasing threats	40% participants viewed social media as risky and challenging for security

Table 2.2: Computer crime Impacts Related Surveys

2.4.1 Individuals

Since an individual forms the most basic unit of society it is important to study the impact of computer crimes on individuals. The amount of losses suffered by individuals in computer crime incidents was 110 billion dollars according to the Cybersecurity 2012 survey (Norton, 2012). The survey also demonstrated that nearly two thirds of adults who go online suffer from one form or another of computer crime, equating to almost $200 per person in loss. The habits of the consumer have a part to play in this, as the same survey revealed that 44% of these adults were not careful about security (such as used unsecured Wi-Fi to access emails). The impact is also aggravated via their use of social networks as 17% of users using social networks suffered from cyberattacks, and three quarters of respondents believed that the cybercriminals are targeting social networks with malicious motives.

Saini et al (2012) suggested that the rising number of incidents of online fraud and scams also affected consumer confidence significantly. Nearly three quarters of respondents had a tendency to refuse a transaction whenever they were asked for their credit card information, for the fear of getting duped or cheated. This is certainly not a good situation considering the growth of e-commerce. It has also been found that many people are not actually aware of the potential damage that can be caused in real life via computer crime, and hence are unable to correlate the two.

2.4.2 Businesses

Businesses and corporations may be categorised into a variety of divisions such as public sector companies, private companies, national government organisations, and Charitable Organisations (Martin & Rice, 2011). Businesses are depending ever more on the Internet not only to provide services to their clients but also their entire business function (Turber & Smiela, 2014). Therefore, it is not surprising that such organisations have been impacted by the growing computer crime phenomenon.

The impact of reduced consumer confidence, as discussed in the previous section, has direct implications for businesses. Moreover, it is somewhat problematic to analyse the exact impact of computer crime on business as a whole since many organisations prefer not to disclose such losses, for fear of loss of reputation (Saini et al, 2012). To gain an idea of the extent of damage such crimes can inflict, the recent TJ Maxx breach could be singled out,

where 45 million credit card numbers were stolen by an employee of the United State Security Service, who was later sentenced to more than 20 years of imprisonment (IGRE, 2012).

As revealed in a study by Ponemon Institute (Ponemon, 2015) the average cost of a data breach to a company was estimated at $3.5 million in the year 2014, a 15% increase from 2013. Studies also suggest that since the trend of employees carrying their own mobile devices in the workplace has risen, posing various security risks and threats. For the organisation, loss of reputation, litigation damages which could be claimed under the data protection regulations of the country in question and the cost of the breach itself may serve to bring down business for a certain period of time (NIST, 2015).

The rising use of the concept of Bring Your Own Device (BYOD) (GovLoop, 2012) to work also blurs the boundaries between what belongs to an individual and what belongs to the organisation. If the same device is used for personal activities and there is a security breach, the chances of sabotage to the company's data is certain. Combined with the previous discussion about impact of computer crimes on individuals, it can be seen that the risk to companies due to this individual factor is also significant.

2.4.3 Impact on Governments/Nations:

A major concern for nations is the fine line between computer crime on one side and cyber warfare and cyber-terrorism on the other, due essentially to the borderless nature of the technologies involved (Awan, 2014). Public authorities, private sector organisations and individual citizens need to recognise this and take the necessary action to ensure a well coordinated response to mitigate the threats associated with computer crimes (European Commission, 2013).

Governmental agencies across the world have recognised that due to the borderless nature of computer crime it is an important area that needs attention and research. For example, the National Cyber Security Programme in the United Kingdom had a budget of £860 million for the period from April 2011 to March 2016 (NAO, 2014). The United States has allocated $200 million for Federal Network Security, $406 million for Network Security Deployment and $70 million for Cybersecurity R&D (Homeland Security, 2014).

2.5 Conclusion

Chapter 2 has considered the global menace of computer crime, where such crimes have stemmed from and how they are evolving. The chapter has given a historical overview of the problem from its birth about 50 years ago, to the current sophisticated state of computer crime. The statistics detailing the rising global numbers of broadband users in both the developed and developing countries, reinforces the scale of the potential reach of computer crime.

Through examining the impact of computer crime at various levels of society, from individuals to organizations and national governments, it can be seen that computer crime and cybercriminals are constantly changing and evolving over time and look to exploit the latest trends and innovations in technology. Currently, focus is mainly on financial exploitation, and crimes impact all Internet users, from home users to large corporations.

Digital forensics is the principal process by which law enforcement and organizations are able to identify illegal activities that could be performed by a digital device. The following chapter focuses upon the digital forensic methodologies and includes an analysis of the accepted procedures to better appreciate what digital forensics is and how it is undertaken.

3 Digital Forensic Methodologies

The chapter aims to present the current state of the art in digital forensic methodologies with a focus on the analysis phase of an investigation. Various methodologies will be compared and contrasted in order to explore their advantages and disadvantages and also areas which require further research. Hence, the need for developing a standardised operational framework cannot be overemphasised.

3.1 Introduction

One of the main tools to fight the rising menace of computer crimes discussed in the previous chapter is that of digital forensics. It must be emphasised that the digital forensics has been used not only in combating computer crimes but also for traditional crimes. For instance, performing digital forensics analysis on the victim's mobiles and laptops has become a routine process in traditional crimes such as murder cases. Such a practice clearly indicates the significant role of digital forensics during the investigation of a traditional crime because it can answer important questions such as who, what, why, how, when and where.

With a history of nearly four and half decades, digital forensics has certainly come a long way (Chouhan, 2014). Most of the techniques used in the past for digital forensics were mainly data recovery tools, which helped to recreate data in order to investigate what had occurred during the incident. Starting out with challenges such as non-documented procedures, heavy reliance on centralised computer systems and a lack of standardised versions of hardware and software, digital forensics experienced various improvements and overcame different challenges. Today, a new set of challenges, such as the growing size of data, encryption and technical complexity pose a number of obstacles to digital forensic investigators (Garfinkel, 2010; Chouhan, 2014).

The aim of any digital forensic examination is to discover relevant facts and evidence that would be admissible in a court of law, so that the prosecution of the suspects can take place in a legally valid manner. It is necessary to ensure that digital evidence is preserved in such a manner that it can be admitted in the relevant courts (Carrier, 2003).

The onus is on investigators to prove to legal experts and in the court of law, normally not well versed in computer techniques and terminology, that a crime has taken place and that the evidence is admissible (Herath et al 2005 Computer Forensics, Information Security and Law: A Case Study). Pollit (2007), also suggested that normally, the evidence in computer crimes is hidden and needs various methods and techniques associated with digital forensic investigations to derive admissible evidence from the examination and analysis.

From the technical perspective, tools and techniques should be available to be used for extracting evidence from raw data and these tools should be flexible to accommodate advancing technologies in digital devices and related fields. Ayers (2009) described these tools in the context of being generationally (i.e. first and second generations) classified based on various factors such as their reliability, efficiency, speed and accuracy.

All the above requirements demand that proper digital forensic methodologies be applied so that the raw data or evidence is extracted, analysed and portrayed in such a manner that it is legally valid. There are several digital forensic methodologies in use for the above to be effective; details of them will be presented and comprehensively discussed in the following section.

3.2 A Review of Digital Forensic Methodologies & Models

It is preferable to have a systematic approach to solving any problem, and the same holds true for digital forensics. The methodologies and models discussed in the chapter form the basis for development of the digital forensic procedures and techniques which aid digital forensic examiners to solve computer assisted crimes. Hence, it is important to track the development of these methodologies in order to understand the path taken by them over time until present.

Table 3.2 presents a comparison of various digital forensic investigation methodologies from 1984 to the current era. Several criteria were developed to aid the comparison. The presence or absence of each of the five main phases in the particular methodology, namely: identification, preservation, collection, analysis and report (denoted by the symbols I, P, C, A and R respectively), are indicated in the following table:

Phase	Identification	Preservation	Collection	Analysis	Report
Key	I	P	C	A	R

Table 3.1: Key symbols for the investigating processes

It must be noted that these five phases have been mapped onto a common platform for ease of readability otherwise the same term may not be used to denote the specific phase or process. For example, the "report" section might be known as "presentation" in some cases, and the "analysis" phase might be noted as "hypothesis". The last column entitled Comments provides a description of the unique characteristics or differentiators of the specific process.

Digital Forensic Investigation Methodologies	**Authors**	**No. Of Phases**	**I**	**P**	**C**	**A**	**R**	**Comments**
Computer Forensic Investigative Process	Pollitt (1984)	4	✓	✓	✓	✓	✓	Divides the process into physical, logical and legal contexts
Forensics Process Model	Ashcroft (2001)	4	✓	✓	✓	✓	✓	Mainly used by investigation agencies
Investigative Process for Digital Forensic Science	Palmer (2001)	6	✓	✓	✓	✓	✓	Based on DFRWS lines
An Abstract Digital Forensics Model	Reith et al (2002)	9	✓	✓	✓	✓	✓	Based on FBI model and enhancement of DFRWS
An Integrated Digital Investigation Process	Carrier and Spafford (2003)	17 phases organised into 5 categories	✓	✓	✓	✓	✓	Maps physical forensics to digital forensics
End-to-End Digital Investigation Process	Stephenson (2003)	9 phases	-	-	✓	✓	-	Possibility to prevent attacks by modelling vulnerability to prevent attacks
The Enhanced Digital Investigation Process	Baryamureeba and Tushabe (2004)	5 major phases	✓	✓	✓	✓	✓	Uses traceback and dynamite to separate physical and digital investigations
The Extended Model of Computer crime Investigations	Ciardhuain (2004)	13 phases	✓	✓	✓	✓	✓	Focuses on management aspects of DFI
An Event-Based Digital Forensic Investigation Framework	Carrier and Spafford (2004)	5 major phases	✓	✓	✓	✓	✓	Views an incident in terms of events, objects and causes
The Lifecycle Model	Harrison (2004)	7 phases	✓	✓	✓	✓	✓	-
The Hierarchical, Objective Based Framework	Beebe and Clark (2004)	6 phases	✓	✓	✓	✓	✓	A multi-tier process and uses Survey, Extract, Examine or SEE in the analysis

Digital Forensic Investigation Methodologies	Authors	No. Of Phases	I	P	C	A	R	Comments
The Investigation Framework	Kohn, Eloff, and Olivier (2006)	3 phases	✓	✓	✓	✓	✓	-
The Forensic Process	Kent, Chevalier, Grance, and Dang(2006)	4 phases		✓	✓	✓	✓	It integrates forensics into the investigation
The Computer Forensics Field Triage Process Model	Rogers, Goldman, Mislan, Wedge, and Debrota (2006)	6 major phases including sub phases	✓	✓	✓	✓	-	Offers real time analysis solutions for use in cases where time is critical
FORZA – Digital Forensics Investigation Framework Incorporating Legal Issues	Ieong (2006)	8 phases	✓	✓	✓	✓	✓	Based on 6 Ws why, what, where, who, how, when
The Common Process Model for Incident Response and Computer Forensics	Freiling and Schwittay (2007)	3 major phases including sub phases	✓	✓	✓	✓	✓	Attempts to combine an incident response and computer forensics, retaining flexibility at the same time
Two-dimensional Evidence Reliability Amplification Process Model	Khatir et al (2008)	5 major phases	✓	✓	✓	✓	✓	Provides a method to increase evidence reliability
Digital Forensics Investigation Procedure Model	Shin (2008)	10	✓	✓	✓	✓	✓	Lays out a detailed procedure for DF investigation
An Extended Model for E-Discovery Operations	Billard (2009)	10	✓	✓	✓	✓	✓	Caters to techno-legal aspects in modern day complex DF situations
A Multi-component View of Digital Forensics	Grobler, Louwrens, and Solms (2010)	3	✓	✓	✓	✓	✓	Divides DF into various components

Table 3.2: Digital Forensic Methodologies Timeline

It may be seen that more recent digital forensic methodologies do not necessarily have more phases, but the general trend is to divide the methodology into more refined phases and sub-phases, with clarity and complexity being added to newer approaches.

The above presented methodologies have their own set of strengths and weaknesses. Many methodologies try to attack the problem from different perspectives and hence have a relevance based on the situation. For example, the methodology of Billard (2009) was designed mainly to handle the challenges of modern day techno legal environments while the

Khatir et al. (2008) methodology stresses evidence reliability. There are certain qualities which each digital methodology should address.

A methodology should define the identification of relevant data. With the rising volumes of data, it is difficult and expensive to keep all the available data for analysis since this would stretch the storage sources (Garfinekel, 2010).

The analysis phase of the methodology should provide a means to reconstruct the overall picture of what might have happened during the computer assisted crime incident, by connecting the dots, whilst being capable of analysing multi-stage attacks (Zhou et al, 2007) and handling multiple levels of abstraction (Peisert, 2007).

3.2.1 The Digital Forensic Research Workshops

Before further discussing the various methodologies and their features, it would be appropriate to take a look at an event which formed an important benchmark in the digital forensics world.

A major breakthrough occurred towards the beginning of the new millennium, when the first DFRWS was established and met in the 2001. It aimed to provide a common platform for the various bodies (e.g. government and civilian agencies, research institutions and companies) to discusses digital forensic related topics. During the first meeting of DFRWS, a model of the basic steps in digital investigation was drawn up. The outlined steps were from the first step of identification to the final step of decision, as shown in the Figure 3.1.

Identification	Preservation	Collection	Examination	Analysis	Presentation	Decision
Event/Crime Detection	Case Management	Preservation	Preservation	Preservation	Documentation	
Resolve Signture	Imaging Technologies	Approved Methods	Traceability	Traceability	Expert Testimony	
Profile Detection	Chain of Custody	Approved Software	Validation Techniques	Statistical	Clarification	
Anomalous Detection	Time Synch.	Approved Hardware	Filtering Techniques	Protocols	Mission Impact Statement	
Complaints		Legal Authority	Pattern Matching	Data Mining	Recommended Countermeasure	
System Monitoring		Lossless Compression	Hidden Data Discovery	Timeline	Statistical Interpretation	
Audit Analysis		Sampling	Hidden Data Extraction	Link		
Etc.		Data Reduction		Spacial		
		Recovery Techniques				

Figure 3.1: Investigative Process decided at DFRWS (Palmer, 2001)

The DFRWS model was considered as a reference model since it formed the foundation on which digital forensics has been developed on a scientific basis. This model was well received within the academic community as it provides a framework for furthering the research in digital forensics (Inikpi et al 2011). It also offers a detailed outline for various digital forensic investigation processes within the division of different phases: identification, preservation, collection, examination, analysis, presentation and decision.

Preservation involves gathering the relevant data. It is useful for the purposes of digital investigations by ensuring that none of the data is lost or tampered with but preserved according to the related procedures that are normally followed in an accredited digital forensic house. The collection phase consisted of reducing the preserved information to relevant data by the use of appropriate hardware and software tools, using the right sampling techniques so that the target data was much more relevant than the preserved data for the digital forensic examination purposes. The examination includes a close scrutiny of the data for finding traces of the crime by searching for things such as hidden data, pattern matches and validating the data. The analysis phase consists of procedures such as statistical techniques and data mining in order to find results. These results are then presented in appropriate formats and in a manner which may be understandable to legal personnel, who then make decisions based on the presented facts and figures. Each of these phases has been subdivided into various components to give a broad outline of the goal which that particular phase was trying to achieve.

DFRWS was an important step in formalising the arena of digital forensics and it recognised three main facts relating to analysis of evidence. The digital sources of information became increasingly complex and were also less understood, adding to the complexity. Secondly, the digital sources were constantly evolving, and lastly, there was a need to constantly monitor and understand the underlying technologies as they developed. This understanding helps to provide a strong basis for further development of digital forensics (Palmer, 2001). It is one of the most well accepted and recognised methodologies from which most others are derived.

3.2.2 Computer Forensic Investigative Process

Pollitt (1995) proposed one of the earliest digital forensic methodologies which aimed to provide a thorough stepwise procedure for dealing with situations involving digital crime. This model suggested that the investigation process consisted of four main steps: acquisition, identification, evaluation and admission (as shown in Figure 3.2).

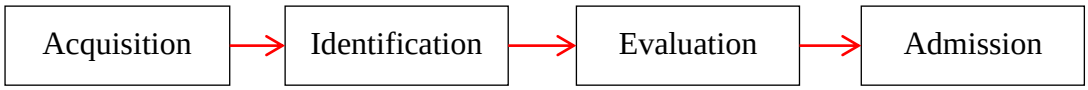

Figure 3.2: Computer Forensic Investigative Process (Pollit, 1995)

Although it does not have an extensive and elaborate coverage, and may seem over simplified in terms of current methodologies, it was nonetheless one of the pioneers of the time, offering direction on how to form a digital forensic investigation. Indeed, all subsequent methodologies include these processes, with either additional stages or more specific details within each process.

In essence, the above model depicted the flow of the forensic process in the offline world. It was argued that while applying this paradigm to the digital forensic world, the evaluation phase should consist of taking care of technical and legal aspects (Pollitt, 1995). The technical aspect can be further categorised into physical and logical based. For the legal point of view, the source of the data and also its admissibility needs to be clear. Hence the process, when applied to the digital world, has the following form as depicted in Figure 3.3.

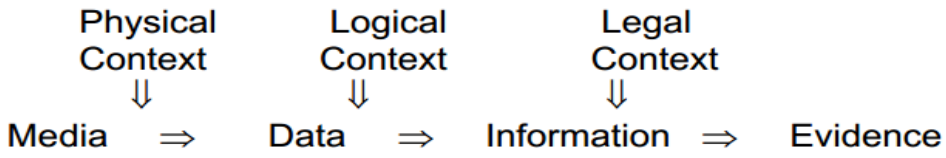

Figure 3.3: Equivalent of Forensic Investigation into Digital World (Pollitt, 1995)

3.2.3 The Forensic Process Model

This model was proposed by the United States National Institute of Justice and its aim was to act as a guide to the first responders on the scene of a digital crime, as they are the ones who play a very important role at the initial stage of the investigation process regarding digital evidence. The Forensic Process Model is also comprised of four stages, namely: collection, examination, analysis and reporting Ashcroft (2001).

The analysis phase in this methodology is quite generic in nature and is not based on any specific technology. It is outlined in detail within the guide and it mainly deals with suggesting types of evidence that could be found on digital devices, the most probable locations, different types of crimes that could be committed using such devices and the nature of possible evidence in each case.

This methodology aims to strike at the core of a digital investigation and could be applied even to future technologies due to its standard nature. In addition, this methodology is seen as an important step in developing more general methodologies with much wider applications, not only for the first responders but also for the complete forensics process.

3.2.4 Abstract Digital Forensic Model

Similar to the Forensic Process Model, this methodology is general in nature. It is an excellent attempt at abstraction by which Reith et al (2002) defined a methodology which was not dependent on any particular technology or type of crimes, and was applicable for cases typically related to the FBI's investigation. This methodology is basically an extension of the DFRWS model; it also consists of nine steps, as outlined in the Figure 3.4 below.

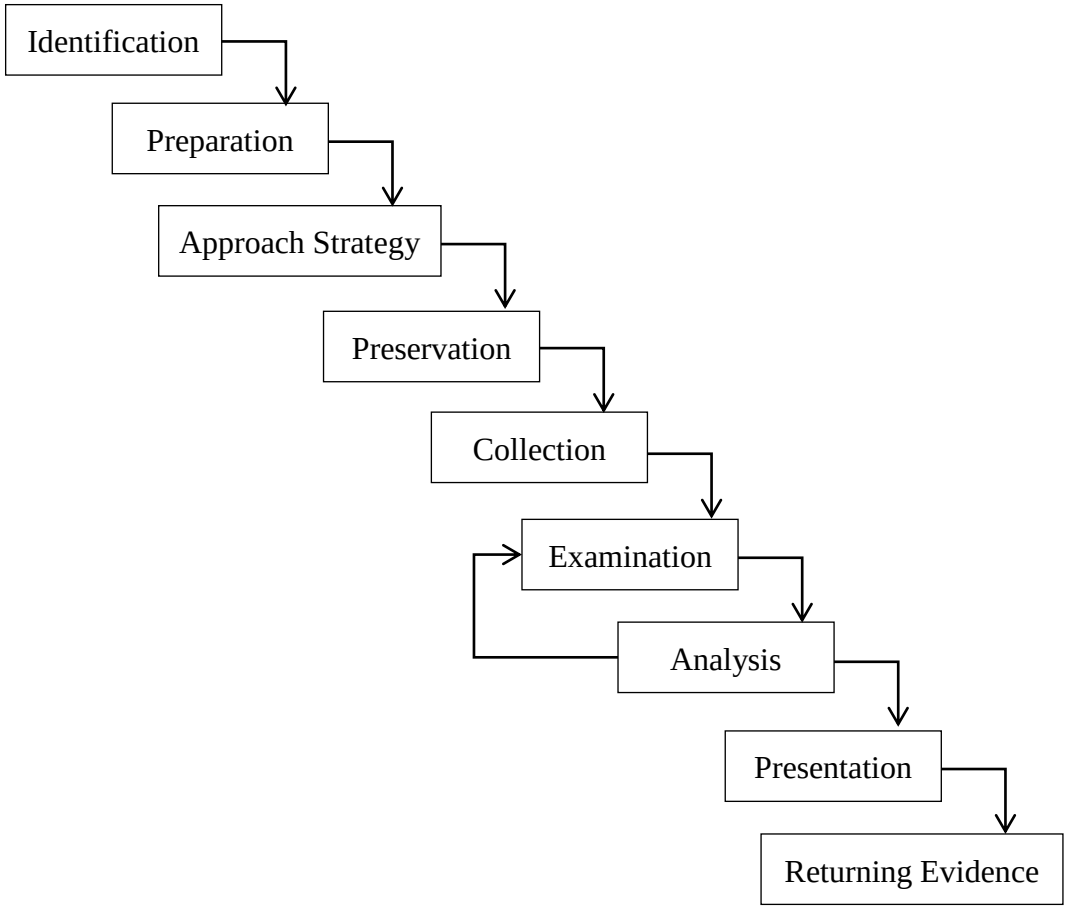

Figure 3.4: Abstract Digital Forensic Model (Reith et al 2002)

As can be seen from the figure, the steps outlined in this methodology are quite similar to those of the DFRWS model but with the addition of the approach strategy and the returning evidence steps. It also provides feedback from the analysis step to the examination phase in order to refine examination criteria, and also has a stage where the evidence is correctly returned after it serves its purpose.

The main contribution of this methodology is that it lays out various stages of the investigation process in abstract units thereby ensuring a common platform for law enforcement agencies and judicial personnel exists which enables them to work in harmony with each other (Reith et al, 2002).

Despite the abstract definition at the highest level, it is still important to define additional sub-procedures based on specific technologies as well. These sub-procedures would help to extend the reach of this methodology at a deeper level to various situations. This is important since at the ground level, the abstractions need to be converted into exact methods and practices (Reith et al, 2002). Since this is a generic methodology, the additional two phases

over and above the DFRWS as explained previously, help to ensure that the abstraction provided in this methodology is applicable to a wide range of digital forensic situations, including computer assisted crimes.

3.2.5 The Integrated Digital Investigation Process Model (IDIP)

Carrier and Spafford (2003) proposed this model in the year 2003 and it contains five phases, which are further subdivided into 17 phases. The sub-phases included in these five phases are clearly presented in Figure 3.5;

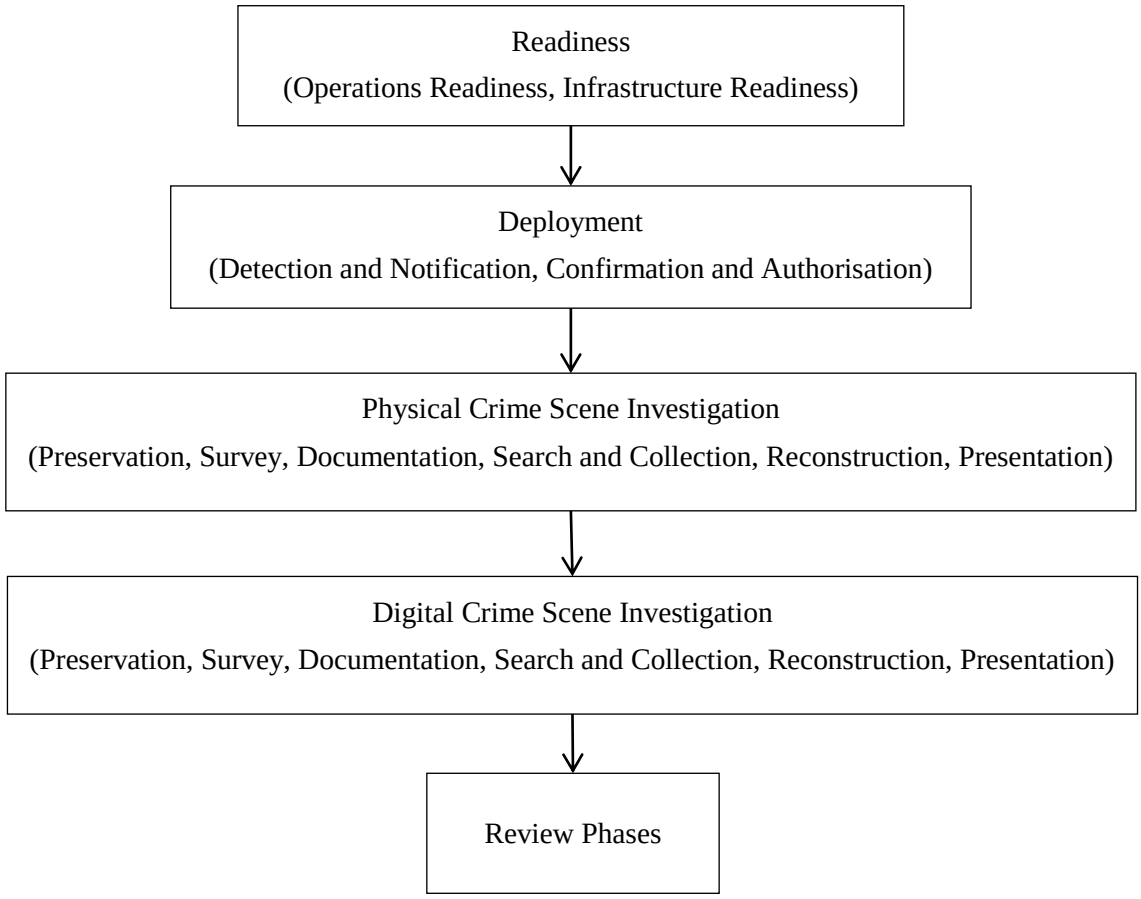

Figure 3.5: Phases of the IDIP Model (Carrier and Spafford, 2003)

The analysis phase consists of two parts for the physical crime scene and digital crime scene investigations, with the digital crime scene able to be considered as an extension of the physical crime scene (Selamat, 2008). Firstly, the physical crime scene analysis is carried out using conventional forensic techniques. Then the digital analysis is performed by using software tools to analyse all sorts of hidden, corrupted files and creating a low level timeline and file log for the purpose of results (Baryamureeba andTushabe, 2004).

The main contribution of this model is that it recognises the time tested practice of physical forensic examination, and draws upon that knowledge bank to introduce processes for this methodology, as it visualises that the challenges in the wake of a digital crime are similar to those faced by physical investigators. Also, this model seeks to bridge the gap between physical and digital investigations by attempting to understand their underlying common aspects and similarities.

3.2.6 Enhanced Digital Investigation Process

With the aim of addressing certain limitations of the IDIP model, Baryamureeba and Tushabe (2006) proposed the Enhanced Digital Investigation Process Model (EDIPM). The main criticism against the IDIP Model is that it defines the deployment phase as distinct and independent of the investigation phases, whilst logically it is not.

As a result, the EDIPM is made of the following five phases (as illustrated in Figure 3.6), with each phase being further subdivided into the appropriate number of sub-phases. The analysis phase is also known as the dynamite phase and it consists of the following sub-phases:

- The physical crime scene is investigated to find out possible clues in the physical realm.
- Digital investigation is carried out in order to create a series of events which possibly would have taken place.
- Reconstruction phase tries to create a complete picture by connecting various pieces of events.
- Communication phase is used to present the entire contents to the appropriate authorities in order for the legal process to take place.

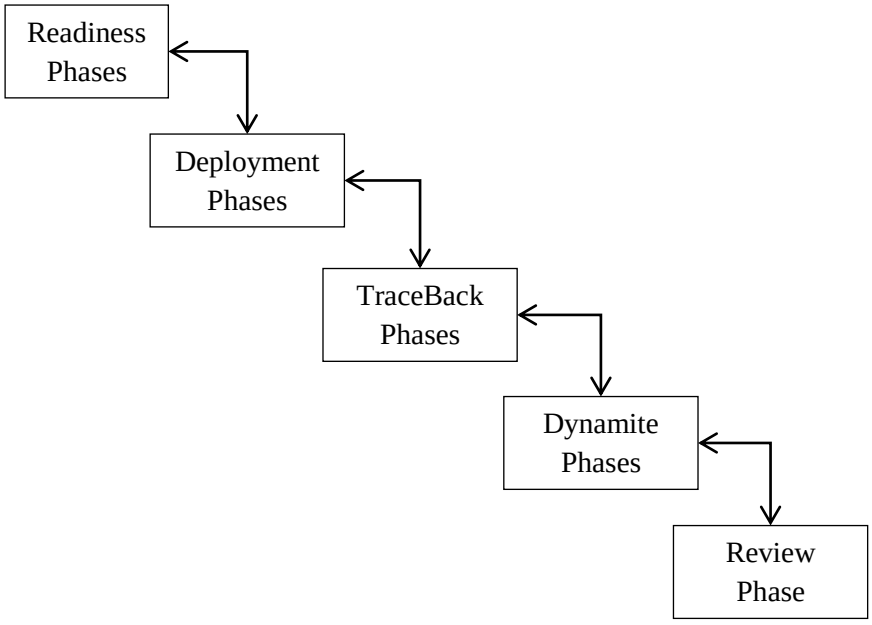

Figure 3.6: 5 Phases of EDIPM (Baryamureeba & Tushabe, 2006)

The basic difference between this methodology and the IDIP model is that the former extends the deployment phase to encompass both the physical and digital investigations in an iterative manner, whereas the latter uses a linear flow between the physical and digital investigation stages. As a result, this methodology is certainly a more practical approach for speeding up the entire investigation process. Also the two-way process ensures that any limitations in the phase are dependent on the previous phase, which results in better outcomes as the output from one phase is acting as the input of the next one. Moreover, this methodology is flexible in terms of dealing with scenarios with various factors such as finding new evidence or data etc.

3.2.7 Extended Model of Computer crime Investigation

Ciardhuáin (2004) suggested another extended model for the investigation of computer crimes in order to overcome flaws that exist within the previous models. These limitations in the previous models include:

- The previous models are not sufficient to cover all aspects of computer crime investigations.
- The lack of provision to capture information flow along the various phases of the digital forensic investigation.
- They mainly focus on evidence collection and analysis, without giving due weight to the other stages.

- They lack the framework to offer a platform for further growth of digital investigative tools and techniques.

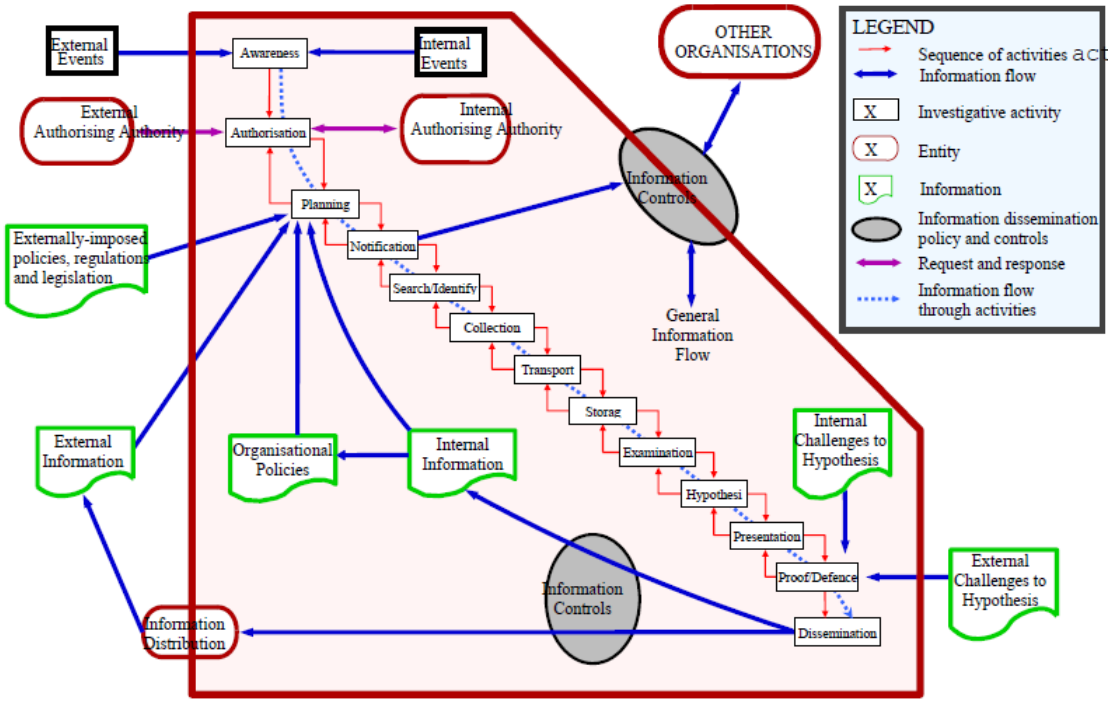

Figure 3.7: Extended Model Flow Diagram (Ciardhuáin, 2004)

As shown in Figure 3.7, this model is based on a waterfall perception of the activities involved in computer crime investigations; and it suggests that it is not always possible for events to flow so smoothly in exactly the same order as depicted, hence the inclusion of backtracking.

The model proposes that a mixture of techniques should be used to analyse the available information. Prior to examination, the data should be repaired in such a manner so that its integrity is preserved. The use of different techniques to measure repair of information often results in a very large volume of data. This requires the use of automated search techniques which would help the digital forensic examiner sift through the enormous data volume.

This model uses the term “hypothesis”, instead of the commonly used term of “analysis”. After the examination phase, as described in Figure 3.7, digital forensic investigators prepare a hypothesis of what they think had occurred. The level of this hypothesis depends on the severity and nature of the case. For example, in the case of a formal legal case involving police authorities, the hypothesis would be formal; while in the case of an IT department of a

company presenting a report to its management, it would be less formal. Here, the concept of backtracking explained above is advantageous because a counter hypothesis is normally presented by the defending party, and in order to prove the fact, the prosecution would most likely need to go back to previous stages, refine the results and gain a better hypothesis.

The main strength of this methodology is that it has sufficient room for information flow between the various stages and phases, which is lacking in the previous methodologies. However, this methodology is rather too generic in nature, and requires lots of details before it can be applied to a specific case.

3.2.8 Case-Relevance Information Investigation

The Case Relevant Information Extraction methodology works on the principle of extracting relevant information pertaining to a specific case (Ruibin and Gaetner 2005). It has been observed that there is a great deal of data in digital forensic investigations which tends to overburden an investigator; as a result, there was a need for an automatic extraction methodology which can select the right amount and type of information required for any particular scenario. Due to this fact, Case Relevant Information Extraction was presented. Figure 3.8 depicts the flow of this methodology and how it helps the investigator save time by reducing the overall work

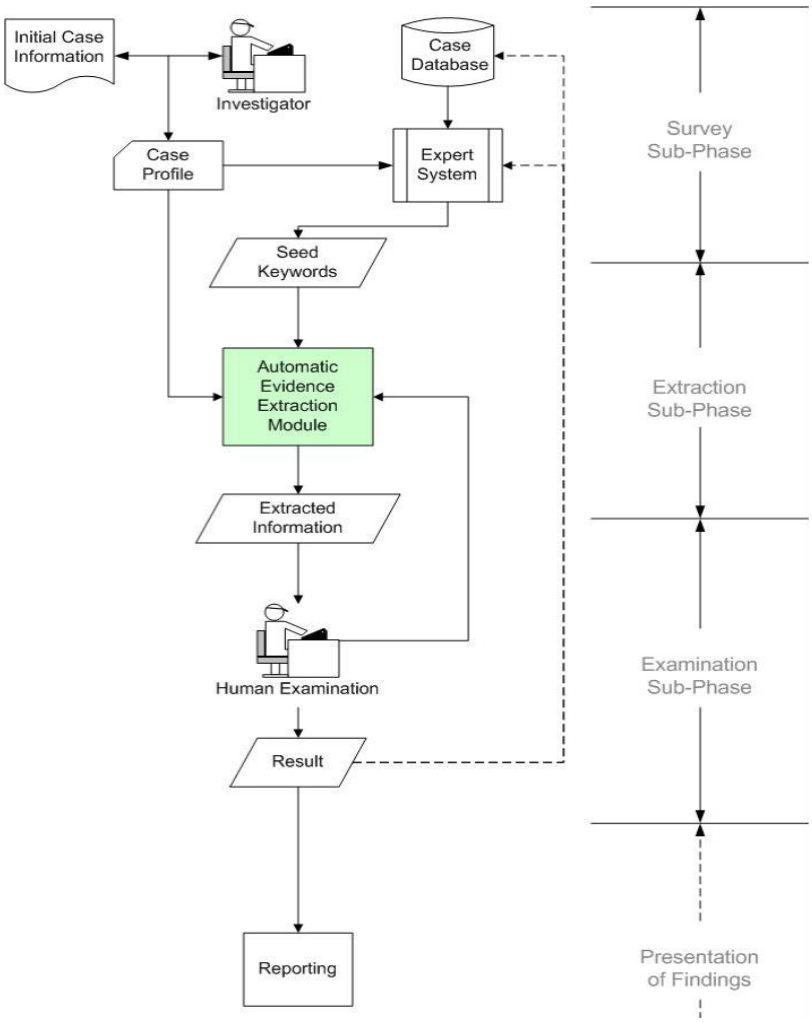

Figure 3.8: Case Relevant Information Extension Sketch (Ruibin and Gaetner 2005)

Despite the fact that human intervention is necessary, as no system can be fully automated, this methodology reduces the workload of investigators and the margin of errors by presenting the right information to be actually investigated. The system proposes the use of automation at second sub-phase level after the initial scrutiny by investigators. The extraction sub-phase utilises automation to extract relevant data by using keywords. This phase is then followed by the human examination phase to check the relevance of the data. Hence, it can be seen that this approach uses an intelligent mix of automation and human intelligence. It must also be noted that even though automation has been proposed as part of many methodologies discussed in this chapter, they have a number of weaknesses such as the lack of proper automated tools, and certain areas of analysis and interpretation which may rely on human intelligence.

Besides consisting of the phases that any methodology would have, this methodology distinguishes between the following phases:

- Survey: at this stage, an experienced investigator is given the task of analysing the available data and preparing an appropriate case profile based on the case background. For example, in the case of pure text data, this could consist of identifying a few phrases or words known as keywords which could act as a base for further searches.
- Extraction: in this phase those keywords are analysed using an automatic system, resulting in additional search terms which ultimately lead to the relevant data in the case.

During the Examination phase, an experienced forensic analyst is required to examine the collected data. It is believed that no matter how effective the automatic extraction toolkit is, it will still not give as accurate results as human beings (Carrier and Spafford 2003). Thus, the final decisions are left to human intelligence since the level of automation is rather simplistic and still places a significant burden upon the examiner.

As a result, this method helps considerably in the analysis phase, since the analysis can only be as reliable as the data available for analysis. Although the final analysis is mainly left to the investigator, this system does deploy an automatic extraction tool which is necessary to ensure that only the most relevant data is available for human investigators to analyse. This ensures that the analysis phase should be able to produce evidence which is admissible in a court of law from various perspectives.

3.2.9 FORZA Framework for Digital Forensic Investigation

FORZA stands for Forensics Zachman framework, and is an attempt to provide a platform in which both legal experts and technocrats have an equal. This framework was proposed by Ieong (2006) and was based upon the observation that the existing frameworks of the time were biased towards the technical aspects of digital forensic investigations.

The ultimate objective of any digital forensic investigation is to find out the perpetrator of the computer assisted crime and bring them to justice, so it is inappropriate to ignore the role of legal personnel in these investigations. Although due to the very nature of the task, the digital forensic investigation does involve a number of technical skills; hence, it is not easy for legal

personnel to get involved in all these complexities. The FORZA framework was meant to overcome this difficulty.

The framework recognises eight different roles which come into play during any forensic digital examination, and defines six questions pertaining to each of these roles. Each role corresponds to a particular layer of investigation (Ieong, 2006).

The analysis layer is used to reconstruct the sequence of events by analysing the data collected during the previous stages. Also, the data analysis phase is considered from both technical and legal perspectives. The model covers eight roles and they are described within the following six questions:

- Why: try to find out the motives behind the case in order to present a detailed digital forensics case.
- What: this consists of the process of reconstructing an event sequence on the basis of recovered evidence in the previous phases.
- Where: this involves finding the whereabouts of the perpetrator and the case from collected data.
- Who: this part seeks to link the events to individuals to find out who is responsible for them.
- How: this involves detecting the involved mechanisms behind the process.
- When: the last stage in this phase is concerned with checking the consistency in events and trying to verify them.

This model seeks to bridge the gap between the technocrats and legal experts by providing them with a common platform; as a result, they can perform their own roles in the best possible manner, without rigidly isolating their specific roles by aiding the technical process and making it easier for non-technical personnel to understand and comprehend.

3.2.10 Forensic Evidence Management System

This framework was presented by Arthur et al (2008) in order to provide a basis for automating the investigation analysis process with the help of crime profiling, and is commonly known as the Forensic Evidence Management System (FEMS). This framework is

based on a component architecture, wherein there are three layers (as illustrated in Figure 3.9), namely: client layer, logic layer and data layer.

The investigation analysis process is automated under this methodology, using the concept of Finite State Automata (FSA) where each state stores information about the past. A hypothesis is made about a scenario and it is then tested using automated procedures with the aid of FSA.

The framework seeks to simplify the analysis stage with the help of automation; however, the functioning of the automata becomes complicated in complex situations. Hence, this framework comes quite close to the requirements of modern day digital investigations but further investigation is required for this framework, especially the role of the FSA.

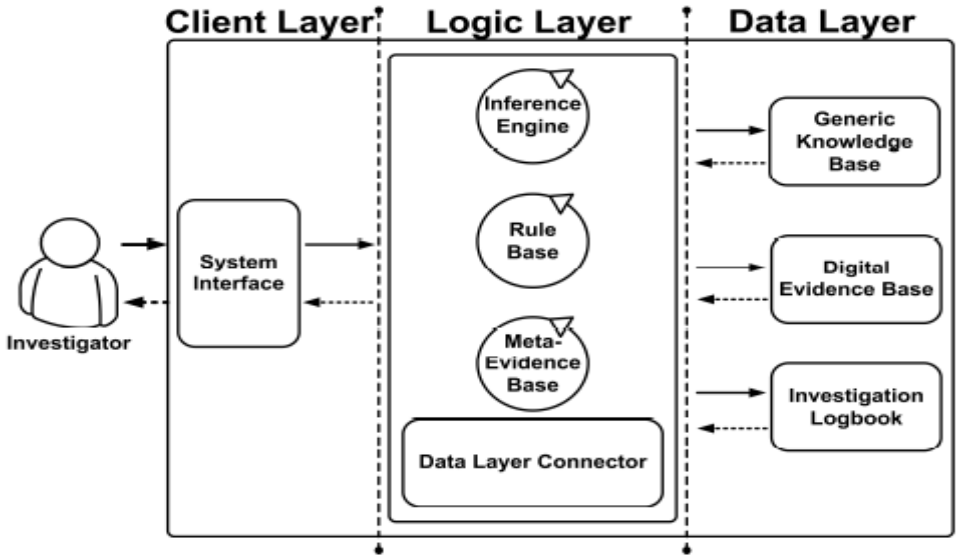

Figure 3.9: FEMS Architecture (Arthur et al, 2008)

3.2.11 Mapping Process of Digital Investigation Frameworks

Selamat et al (2008) proposed a mapping process, whereby all the main existing digital methodologies of the time were mapped by grouping similar activities that produced similar outputs in different phases of appropriate nomenclature. Firstly, the various processes in different existing frameworks were identified; then the investigation process was dissected into five phases, namely Preparation, Collection & Preservation, Examination & Analysis, Presentation & Reporting, and Disseminating. The final step consisted of mapping the various processes and phases of the different methodologies to the above phases. This

methodology leads to the simplification of various existing frameworks which were complex in nature, by eliminating redundant or duplicate steps.

3.2.12 Digital Forensic Model Based On Malaysian Investigation Process

With the aim of providing a better information flow and highlighting important issues (such as chain of custody), Perumal (2009) proposed the digital forensics model based on the Malaysian investigation process. The previous models mainly focused on the processing of digital information, whereas the acquisition of fragile digital evidence also needs attention. The proposed model consists of seven different stages ranging from planning to a diffusion stage (as illustrated in Figure 3.10)

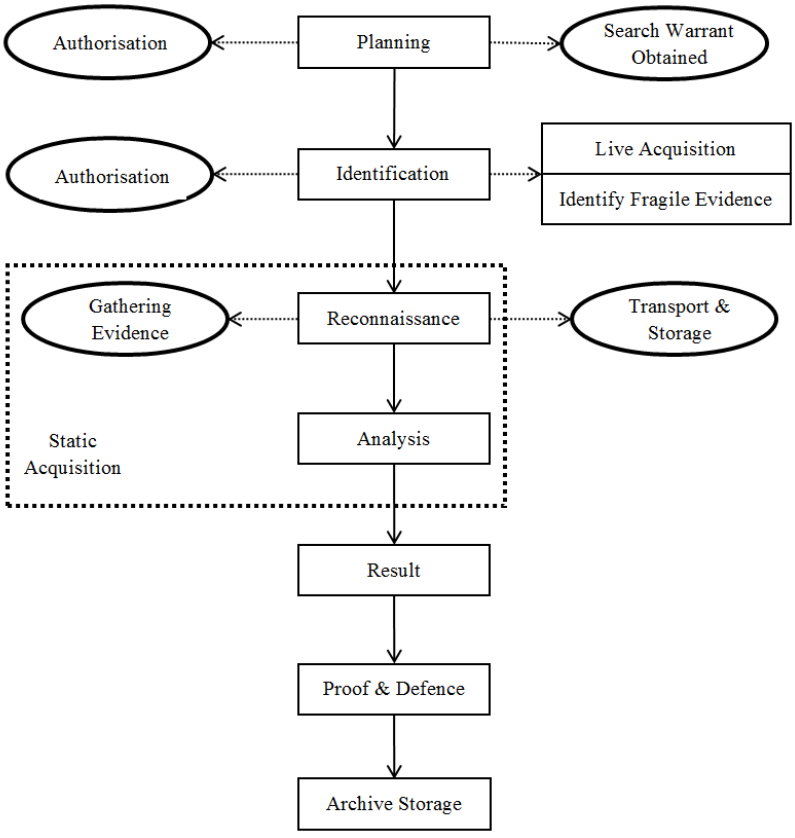

Figure 3.10: Malaysian Investigation Process Model Information Flow (Perumal, 2009)

The analysis stage contains a combination of live acquisition and static acquisition to gather evidence; by utilising a wide range of investigating tools, the acquired data is analysed to enable a consistent and holistic perspective to be observed by investigators. Normally it is seen that individual items do not have the strength to stand as admissible evidence. Yet they play an important and useful role in completing the chain within the investigation as an item could point towards another item which in turn could be an important piece of evidence.

The main advantage of the model is that it claims to overcome the deficiency of a detailed information flow in previously proposed models. However, the model focused mainly on memory data acquisition and data mining techniques. Nonetheless, this methodology is certainly an improvement in comparison with other existing models at that time; arguably more work needs to be focused on devising more effective ways of data acquisition, research and analysis.

3.2.13 Computer crime Execution Stack

The computer crime execution stack was proposed by Hunton (2011b), aiming to visualise the main features in a typical computer crime scenario. The various components of the computer crime stack are shown in Figure 3.11.

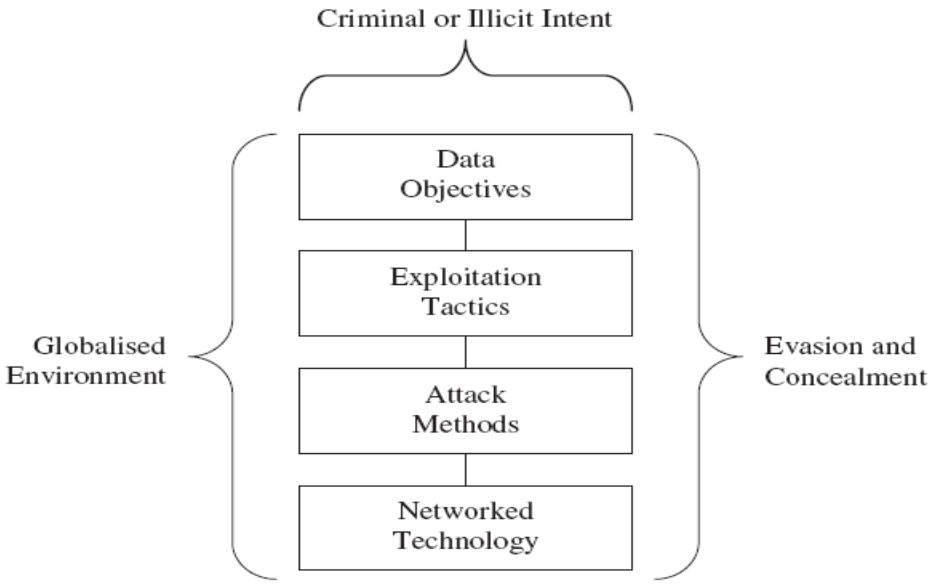

Figure 3.11: Computer crime Investigation Stack (Hunton, 2011b)

The researcher claims that this model can be used to provide substantial support during the analysis phase of the digital forensic investigation, as outlined in the ACPO Core Investigative Doctrine. Since computer assisted crime is becoming ever more complex, various networking technologies could be used in a single incident (ACPO, 2012). These need to be broken down into individual components in order to perform a complete analysis. Since the nature of digital evidence is normally very fragile and delicate, it is possible that the incorrect order of assessment of the various devices and technologies used in the computer

assisted crime may lead to deterioration or destruction of the evidence. Hence, this framework seeks to prioritise tasks based on their technical classification and is a useful platform for further research and development in the increasingly complex environment of computer assisted crime scenarios over time.

3.2.14 The Systematic Digital Forensic Investigation Model

This model was developed by Agarwal et al (2011) to overcome the perceived drawbacks in the previous models, which faced several challenges due to various factors including (but not limited to) rising volumes of data and diverse media types.

In essence, this is an attempt to systematically provide a path for digital forensic analysts to help various stakeholders to set up efficient policies and system procedures. In total, the Systematic Digital Forensic Investigation Model (SDFIM) consists of eleven phases (as illustrated in Figure 3.12).

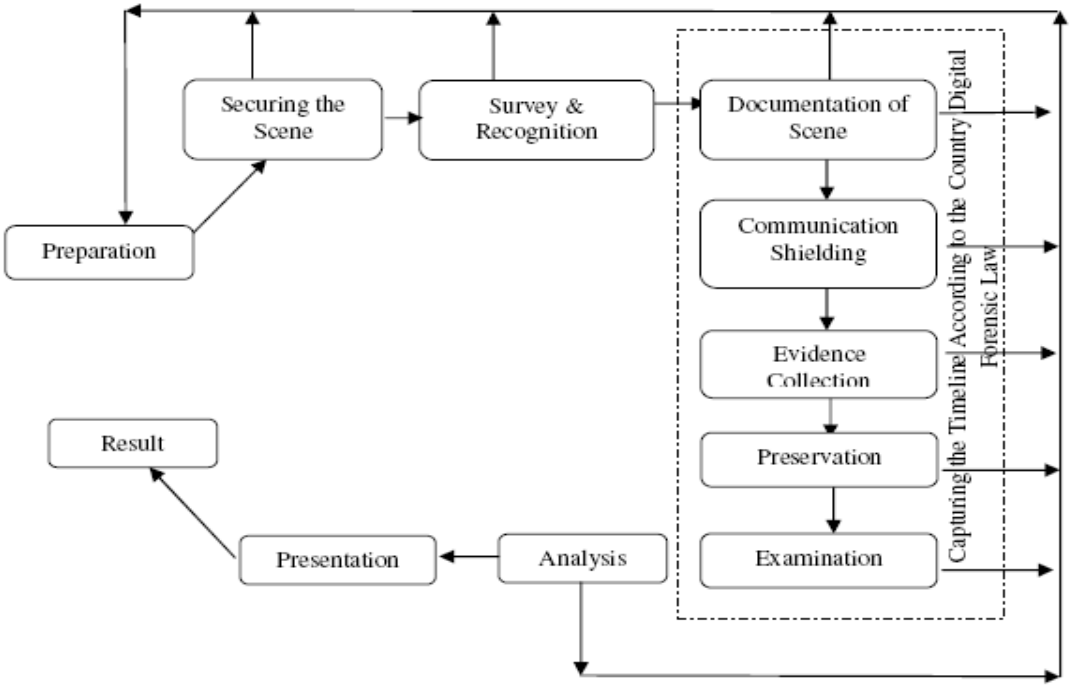

Figure 3.12: Phases of SRDFIM Model (Agarwal et al, 2011)

Since the Examination phase is of particular importance to this research, this phase is described in detail. The most important component of the examination is to convert the plethora of data into manageable chunks of information using techniques such as pattern matching, keywords search and filtering. Apart from sorting out the visible information, one

of the main components of this phase is to dig for or uncover hidden information. It also involves the logging of all the actions and uses techniques like hashing to ensure integrity of the records. Moreover, various toolkits can be used during the examination phase by digital forensic examiners, which determines the quality of output and the time spent on such efforts.

In the analysis phase, the SDFIM ensures that the data extracted during the examination phase is subjected to a thorough technical review which will help to ensure that the various segments of collected information are connected in some form, leading to a bigger picture of the incident. This will help to identify the perpetrator and the production of evidence necessary for a court of law.

3.3 Discussion

From the preceding sections, it demonstrates that various methodologies were developed over time to assist digital forensic investigations. The majority of these methodologies were devised to overcome limitations within the existing methods of the time and to expand the horizon of digital forensics.

Since examination and analysis are the core aspects of any digital forensic investigation, they were the main discussion focus of this chapter. As emphasised during the individual analysis of the various methodologies above, it is also apparent that this does not mean neglecting other important aspects such as data acquisition and so on. Some of the major points which arise out of the research completed in this chapter include the following:

Automation: it has been noted that due to several factors, especially rising computer assisted crime rates and increasing volumes of data which require sifting through by digital forensic examiners, automation is acquiring the status of a necessity rather than a luxury for the examiners. Yet there are issues of thorough automation techniques which actually tend to at least partially replace the human factor. The current efforts at automation have been few and far between, and there is certainly a lack of proper tools which could be of high value to the examiners, helping them to focus more on the actual investigations rather than routine procedures and tasks.

Usability: any methodology, in spite of its level of advancement from the research or academic viewpoint, should have practical implications for the digital forensic examiner at ground level. This will ensure that it actually helps personnel who deal with computer crime incidents on a day to day basis, and in a manner which is not too difficult to comprehend or implement. The usability factor is certainly one of the vital links which relates the research done in labs to the applications in the field.

Reliability and Acceptance: these two factors are very closely related, and normally a methodology which is reliable will have better acceptability rates. Better adoption rates also ensure that standardising the procedures is easier which results in having a common framework globally, despite the boundaries of nation states. This is also an important requirement in scenarios relating to cyberspace, where boundaries do not exist at all.

Adaptation to evolving threats: it is critical that any methodology should have the ability to adapt evolving threats; without having this factor in mind, the methodology could become exhausted very quickly if new threats appear. Apart from the early generic models (e.g. the DFRWS model and the Forensic Process model), majority of recent works claimed that their proposed methodology was designed to overcome limitations of existing methods (e.g. dealing with new threats) and offer new functionalities. This demonstrates that new proposal does take the adaptation to evolving threats into considerations to deal with the complex nature of the digital forensic investigation and fast evolution of threats within the IT domain. As a result, any novel methodology should consider potential new threats.

Amongst these areas, the usability, reliability, acceptance and adaptation to evolving threats are the basic requirements for any forensic methodologies; in other words, a new methodology or tool will not be accepted by the digital forensic domain if any of those four were missing. In comparison, the automation offers additional functionality that could be used to save investigator's time by speeding up the analysis process in a more accurate way; hence, the image data can be processed in a timely fashion. Also, as mentioned earlier, little research or work has been done on this topic. As a result, this result will focus upon the investigation on an automation method that can be used by forensic investigators to analyse a case image in an automatic manner.

3.4 Conclusion

It may be perceived that the digital forensic methodologies have been improving over time. These methodologies provide the framework on which digital forensic analysis rests.

It can be concluded from the above discussion that starting from very basic beginnings nearly three decades ago in 1984 until 2011, digital forensic methodologies have matured significantly, covering previously uncovered areas and trying to examine as much ground as possible. Indeed, the sequential study of the methodologies clearly indicates that there has been substantial progress with each subsequent methodology, building on the previous ones and adding something new to the already existing knowledge base. As a result, the methodologies have been improving from various aspects, adding to the reliability, usability and consequential acceptance within the digital forensic circles.

Within these methodologies, the examination and analysis process is mainly carried out by human investigators; such a process poses increasing challenges due to the large amount of data presented within forensic images. As a result, there is a strong need to develop automation tools which not only help to save forensic examiners' time but need to be also reliable and widely accepted. To this end, next Chapter will focus upon examining existing tools that are used for conducting digital forensic investigations, and to what extend the automation is utilised for the examination process.

4 An Analysis of Digital Forensic Tools

The previous chapter undertook a detailed investigation examining various methodologies utilised within the digital forensic domain in a chronological order. In this chapter, an evaluation is undertaken of current open-source and future investigative tools. This will involve presenting their functionalities and discussing their limitations; the triage concept and the use of automation will also be described.

4.1 Introduction

This chapter analyses the capabilities and limitations of first generation digital forensic tools, and highlights and compares the capabilities and features of the researched tools. It is envisaged that the review outcome will help to determine the capabilities of future digital forensic tools required to respond to the constant increasing rate of computer crimes. Therefore, this chapter mainly concentrates on the most common open-source digital forensic tools in order to create an improved method and technique that shall continue to be fit for the purpose of digital forensic investigations.

4.2 Requirements of Second Generation Tools

Carrier (2002a) presented a list of the desired features of digital forensic tools, including usability, comprehensiveness, accuracy, and deterministic and verifiable results. Ayers (2009) further defined the existing digital investigation tools as belonging to the first generation because they are mostly helpful in the process of evidence recovery rather than actual case examination and analysis; this means that the tools are mainly used to process the data; but the main examination and analysis on the data is carried out by a human investigator. Current tools hardly offer any decision capabilities to help the forensic investigator during the case analysis phase. Therefore, future tools should be equipped with such capability to remove parts of the investigator's workload.

Another weakness of current tools is the lack of re-usability of data and knowledge gained from previous investigations. It is widely understood that previous experience could be valuable for solving current cases, however most forensic examiners agree that the use of previously gathered data is relatively limited (Horsman et al 2012). Therefore, future trends

need to be focused towards developing a knowledge base which can be used by the entire digital forensic examination community during investigations.

This could only be achieved if the processes of digital forensics are standardised and have a common benchmark across the various countries and regions (Horsman et al 2012). This would also help to ensure that uniformity is achieved and leads to a harmonious development of digital forensic frontiers worldwide. However, even if such a common knowledge database existed, it would not be compatible across the investigations carried out by different individuals because there is a lack of formal investigation processes.

Ayers (2009) further suggested that the following features should be desirable in second generation digital forensic tools if they had the ability to provide sufficient support in terms of rising data and speeds. These factors are discussed below.

- Higher speeds: second generation tools certainly need to have much faster speed and this can be achieved using a combination of various approaches including (but not limited to) supercomputing, grid computing and parallel computing.
- Higher accuracy: the use of current digital forensic tools can achieve a good level of accuracy. Due to the nature of digital forensic examination and its ultimate aim to convict suspected individuals, it is critical that the highest accuracy can be achieved by using these tools.
- Higher completeness: this means that the tool is capable of finding as much evidence as possible from a given set of data.
- Higher auditability: this is an important parameter due to its role in the legal aspects of digital forensic investigation. Auditable results mean that they can be cross verified making them solid legal evidence.
- Higher automation levels: automation is one of the most important aspects of digital forensic tools as it can be used to save investigator's time and effort. As a result, the backlog of computer crime cases could be reduced.
- Faster I/O: most operations related to a digital forensic examination are related to input and output operations (i.e. data is transferred across devices and media during various stages). For instance, the speed of data processing has a significant impact upon the analysis time: the quicker the data processing, the shorter the analysis time. This cannot be achieved by simply improving data storage devices but also needs the improvement of disk storage formats as well. Indeed, first generation tools have the

severe limitation of using an almost identical approach to disk storage formats despite increasing storage capacities; therefore, this limitation should be addressed in second generation digital forensic tools.

- Higher comprehension levels: future tools should be capable of providing the information at higher abstraction levels in a more humanly comprehensible format which anyone could interpret, not just technocrats. This is important since the area of digital forensics encompasses experts from a variety of sectors. Legal personnel such as lawyers and judges play one of the most significant roles during the actual prosecution and conviction stages, which are ultimately the goals of any digital forensic investigation.

Ayers (2009) proposed possible solutions to such problems and ways to develop the second generation systems, which could potentially address these requirements and overcome the limitations of the first generation forensic investigation tools.

Several options have been suggested to increase the processing power of analysis systems, such as the use of Beowulf clusters, IBM Bluegene clusters or the use of Grid computing (Ayers, 2009). Techniques such as the use of data clusters for storage could solve storage limitation problems, and the use of more reliable software processes would consequently deliver greater systems reliability.

Hence, it may be argued that various attempts are being made to overcome the shortcomings of the current first generation tools, in order to develop faster and more powerful and reliable second generation forensic investigation tools. This is a continuous process that needs to be undertaken on a joint basis amongst various researchers, agencies and authorities; therefore a uniform level of growth in various sectors and regions of the world can be achieved.

This practice will only ensure that progress in digital forensics investigation is enhanced and better developed. It has also been reiterated several times that building scalable open source tools may be a better option than using expensive proprietary tools (Roussev 2011). This will not only give wider access to people from different areas who can contribute their knowledge to develop improved tools and approaches but also provide a wide testing ground to validate the capabilities of the improved tool sets. It would certainly be beneficial for the digital forensics community as a whole, although Carrier (2002b) suggested that extraction tools

could be open source and the presentation tools can be closed sourced with a published design for the purposes of legal admissibility.

The courts are normally concerned with two main aspects of the results produced by digital forensic tools, namely reliability and privacy protection of the general public (Adams, 2008). The concept of reliability is closely linked to the performance that a tool can provide in terms of accuracy and usefulness to an examiner on the ground.

There are various functionalities which an ideal digital forensic tool should be able to fulfil, including hashing (Roussev et al, 2006), data carving (Craiger, 2010), decryption (Casey, 2002) and Steganography analysis (Kessler, 2007). Although many tools seem to have several common capabilities and are well suited to a specific application or set of applications, there is still a need to develop tools which can be used for dealing with a much wider range of tasks.

4.3 The Triage Concept

Having considered some of the different types of digital forensic tools which are currently being used, it would be appropriate to look into future trends too. One of the most important trends seen in the tools of the future is the use of the triage concept. Triage refers to the concept of sorting out evidence based upon its priority and relevance to the given case. The concept of triage has already been in use in various domains, such as in the medical field in order to find the priority of patients for treatment (Lim et al 2012). It is used whenever limited resources are available for relatively larger demands.

Over time, the figures relating to computer crime have been on the rise due to increasing volumes of data (Ayers, 2009), variety of electronic devices (Hunton, 2009), growing numbers of Internet users (Hunton, 2009) and many other factors. Digital forensic investigation fulfils the necessary prerequisites for the application of triage as digital crimes are rising at a fast pace while the available resources in terms of trained manpower and time are limited. Therefore, digital evidence has to be prioritised during investigations (Hunton, 2011).

One important model based on the triage concept is known as the Computer Forensics Field Process Triage Model (CFFPTM), and it has the following objectives (Rogers et al, 2006):

- To find usable evidence as quickly as possible;
- To find the victim which is at high risk;
- To give a direction to the on-going investigation;
- To find the level of danger that the suspect poses to society.

All these factors are addressed in six phases, as follows;

- Planning: a quick and thorough plan of assessing the incident is prepared, indicating which evidence is required.
- Triage: various factors are listed based upon their priority and the most volatile items are dealt with first, and vice versa.
- Usage: after storage media has been prioritised in terms of triage, the link between user profiles and the actual device usage is established.
- Timeline: a chronological sequence of events is established, helping to identify various points in the crime chain.
- Internet: the artefacts associated with the Internet activity are examined, including the use of browsers, emails and messengers.
- Evidence: the evidence found is related to the given case and specific circumstances.

The use of triage is very important for speeding up the process of digital forensic examination and reducing backlog (Horsman et al 2011). Nonetheless, triage tools also have certain limitations which need to be addressed in future, and this can be mainly achieved through the process of automation (Horsman et al 2011) making existing triage tools suitable for a wider range of computer crime scenarios.

4.4 Digital Forensic Automation

As discussed in the previous section, the field of digital forensics is experiencing new challenges due to the expanding use of the Internet, rising volumes of data and information, and varied devices being used by people (Hunton, 2009). As a consequence, digital forensic examiners would have to provide more effort to assimilate this quantity of information and focus less on the actual analysis and investigation, making the situation even more difficult. Therefore, efforts should be directed towards making automated processes and tools which

can aid the digital forensic investigators during an investigation, enabling them to cope more efficiently with the rising numbers of crimes.

Despite many tools have been developed (both commercial and open source) and are currently being used by digital forensic investigators and researcher on a daily basis, such as EnCase (Guidance Software, 2015), Forensics Toolkit (FTK) (AccessData, 2015), P2 Commander (Paraben Corporation, 2015), Autopsy (Carrier, 2015), HELIX3 (e-fense, 2014), Free Hex Editor Neo (HHD Software, 2015) and Bulk Extractor (Garfinkel, 2013), they are still mainly used for presenting and organising the data with a forensic image and little automation is offered during the case examination and analysis process. As a result, some of the prominent automation efforts will be considered and assessed to determine how effective they have been in providing some relief to the investigators. Details of these tools and associated analysis upon them are presented the following subsections.

4.4.1 Case-Base Reasoning

Researchers such as Casey and Friedberg (2006) believe that it is not easy to fully automate the entire digital forensic examination process as the examination process is very complicated (i.e. case dependent) and trained human examiners are irreplaceable by current technology. Therefore, automation can mainly be applied to routine tasks rather than tasks requiring intelligent reasoning (i.e. what human beings are capable of doing). However, questions remain unanswered regarding the level to which automation may be applied within a digital forensic investigation. One of the approaches of automation is the use of the Case Based Reasoning (CBR) technique. In simple terms, the CBR concept attempts to provide solutions to the problem based on its knowledge base (information obtained from previous investigations).

The CBR concept is based on the four Rs, namely: Retrieve, Reuse, Revise and Retain, and these four stages are described as follows (Horsman et al 2011).

- In the retrieve stage, the system decides which case best suits the current situation; so the information from that case can be used.
- In the second step, the knowledge on the existing case is reused and applied to the current one in order to obtain an accurate solution.
- Then the proposed case is adapted based upon the given information.
- Finally the case is kept or retained for future reference in the same knowledge base, completing the CBR process.

Hence, the knowledge base plays a vital role in the CBR technique, and the automation of the examination process can only be achieved if sufficient relevant information exists in the knowledge base. This can only be ensured by continuously updating the knowledge base and also filtering out irrelevant information to keep the base in a maintainable size.

The CBR approach heavily depends on the information stored in the knowledge base, which is stored in a fragmented form rather than a complete solution. One of the main limitations is that an exact match (or a likely solution) between a given case and previously existing entries may not be present when the CBR is applied to the case. As a result, this technique will not be suitable for all situations. Further research directions for the CBR system could be the development of a systematic approach for the knowledge base and frequent validation of its output to ensure it is the same or as close as possible to a human forensic expert's decision.

Another important component of the CBR is Case Profiling that is based on the concept of keeping crime profiles wherein a criminal could be identified according to previous crime patterns. For instance, traditional forensic investigations such as those by Baumgartner used profiling to pin point the suspect from a database based on the links between various artefacts (Horsman et al 2011). By applying the Case Profiling technique CBR can utilise the unique patterns of previously recorded digital crimes for understanding and identifying the given case in a much quicker manner and with less effort.

At a higher level, each computer crime has a set of features which are unique to that crime and these can be profiled to cross check for possible matches in any future crimes. If any correlation is found between these features and the data of the investigation, it may be a useful indicator of the possible suspect.

The main reason for the lack of use of case profiling is that very few researchers create profiles that would be used in future. This lack of profile information is causing slow acceptance of case profiling within the digital forensic circle. A good attempt at profiling was made through the Forensic Evidence Management System (FEMS), which can be used to create profile evidence on the target drive and subsequently used by constantly updating itself and learning on a case by case basis (Arthur et al, 2008).

A similar approach to profiling was adopted by Corney, focusing on user profile storage on various operating systems (e.g. Windows XP) (Horsman et al 2011). By keeping a track record of the activities related to the user account as well as outside its confines, the forensic investigator can find a wide range of suspect behaviours.

Overall, case profiling is a useful technique that can be used to quickly identify crime patterns in a crime scene and relate them to previously committed crimes to arrive at convincing conclusions.

4.4.2 XLIVE

This is a further attempt to provide a framework for live forensic analysis, specifically targeted at Windows operating system based machines. This is known as XLIVE and is an XML based framework for the data collection of live digital forensic investigations (Lee et al 2010). XLIVE is useful in analysing large amounts of data in real time which would be otherwise too difficult and time consuming to analyse posthumously. The XLIVE framework is based on three main building blocks which are described below (Lee et al 2010):

- The first basic building block focuses on circumstances when it is very difficult for the forensic examiner to take a live digital primary copy of the evidence or systems under consideration. This difficulty could arise from a variety of factors such as the massive amount of data, lack of time for data collection, and in case of mission critical systems, the inability to simply switch them off (to avoid the disruption of important services).
- The second building block concentrates on the concept of automation and how to choose the most appropriate approach for the case, based on its crime profile.
- The final step is to gather and present relevant data in the popular XML format. By using the XML format, data is presented and organised in a tree format that is universally accessible.

Digital Forensics XML (DFXML) is a language developed by Garfinkel (2012), aiming to add the feature of composability in forensic tools; this can be achieved by creating a common platform where the majority of the available digital forensic investigation tools are capable of importing and exporting data. The lack of such a common language format is one of the biggest hurdles in current digital forensic investigation scenarios (Garfinkel, 2012).

DFXML seeks to provide a solution for such data transportability by providing the following features (Garfinkel, 2012):

- The researcher tried to provide an easy to use platform so that it is not difficult to implement.
- Attempts were made in a way that the existing file formats do not need to be replaced completely but should be complemented.
- The solution should be scalable so that it can fit the extended requirements over the required time.
- The different parameters of data, such as source identification, are available at one place using XML elements of the appropriate type.

There are tools available that use this language to convert disk images into data in the DFXML format which can then be used for digital forensic investigations.

4.4.3 Automatic Windows Log

This forensic tool was developed in an attempt to automatically create a windows log from the Windows NT based operating systems, including Windows XP and 2003 (Murphey, 2007); also this tool creates the log in a single step without manual intervention (i.e. in an automated fashion). Moreover, it is a useful methodology to create a log of events when direct information is not available such as files without time stamps. This automation process has four stages, each of which is described briefly as follows (Murphey, 2007):

- Recovery – at this stage the files are required to be recovered with the solution being to find all files in unallocated sectors which are still intact. Various types of tools can be used at this stage, such as Scalpel.
- Repair – at this stage recovered logs are repaired to address any log corruptions.
- Validation – in this stage, collected and repaired logs are verified and validated by using credible tools, such as the LogParser from Windows.
- Collation – this is the final step where all recovered data is collected and combined to form specified series of events which can then be used as forensic evidence for further analysis and investigation.

Hence it can be seen that this automated windows recovery process is essentially an automated attempt to recover all forensic evidence without any manual intervention from a human investigator.

4.4.4 Automation Based on B – Method

As cyber criminals become more sophisticated, they tend to erase any trace of their presence by altering files and figures which contain important information regarding the activities carried out on the system from local or remote locations (Gladyshev and Enbacka 2007). Hence, simple log analysis may not be appropriate in such a case to find out what exactly took place.

Gladyshev and Enbacka (2007) provided an automated method for tracing such irregularities and inconsistencies where deliberate attempts were made to hide traces by tampering with normal log files. The basic principle underlying this automation attempt is that multiple data structures are involved in logging various activities and the perpetrator would most likely leave some kind of trace. As a result, the inconsistency can be useful to pinpoint the problems that exist regarding that data or log alterations.

The approach is based upon the well-known B-method. The B-method is a well-known and widely accepted formal method for system development, and has been used for creating safety critical computer systems which hold certain safety properties under all circumstances (Clearsy, 2016). Therefore, Gladyshev and Enbacka's (2007) method has the advantage of undertaking a rigorous analysis of the inconsistencies in the use of the B method; and any found irregularities are converted into plain SQL statements which are then handled by using a specially developed tool for this purpose. The drawback of this process is the insufficient quantity of information in the domain of using formal methods for digital evidence analysis. This should therefore be addressed in future developments.

4.4.5 FACE

The Framework for Automatic Evidence Recovery and Correlation (FACE) is a proficient attempt by Case et al (2008) where the researchers developed a solid automation framework and also presented a tool called *"ramparser"* for automation in Linux based systems. Ramparser can be used to create a memory dump of live systems; the dump contains valuable information relating to network sockets, connections and running processes and applications.

The output is then correlated with other evidence within the case to form an overall picture using the FACE framework.

As demonstrated in Figure 4.1, FACE has five main data views (Case et al, 2008):

- Users: presents a brief or detailed overview of the users in terms of their user IDs, time of last login, and the processes used by that person etc.
- Group: illustrates an overview of the groups in terms of group ID and related information.
- Processes: list the processes in terms of process ID, hex dump and all the related information.
- Filesystem: gives an overview of the entire file system structure.
- Network capture: offers a detailed view of the packets and their associated information.

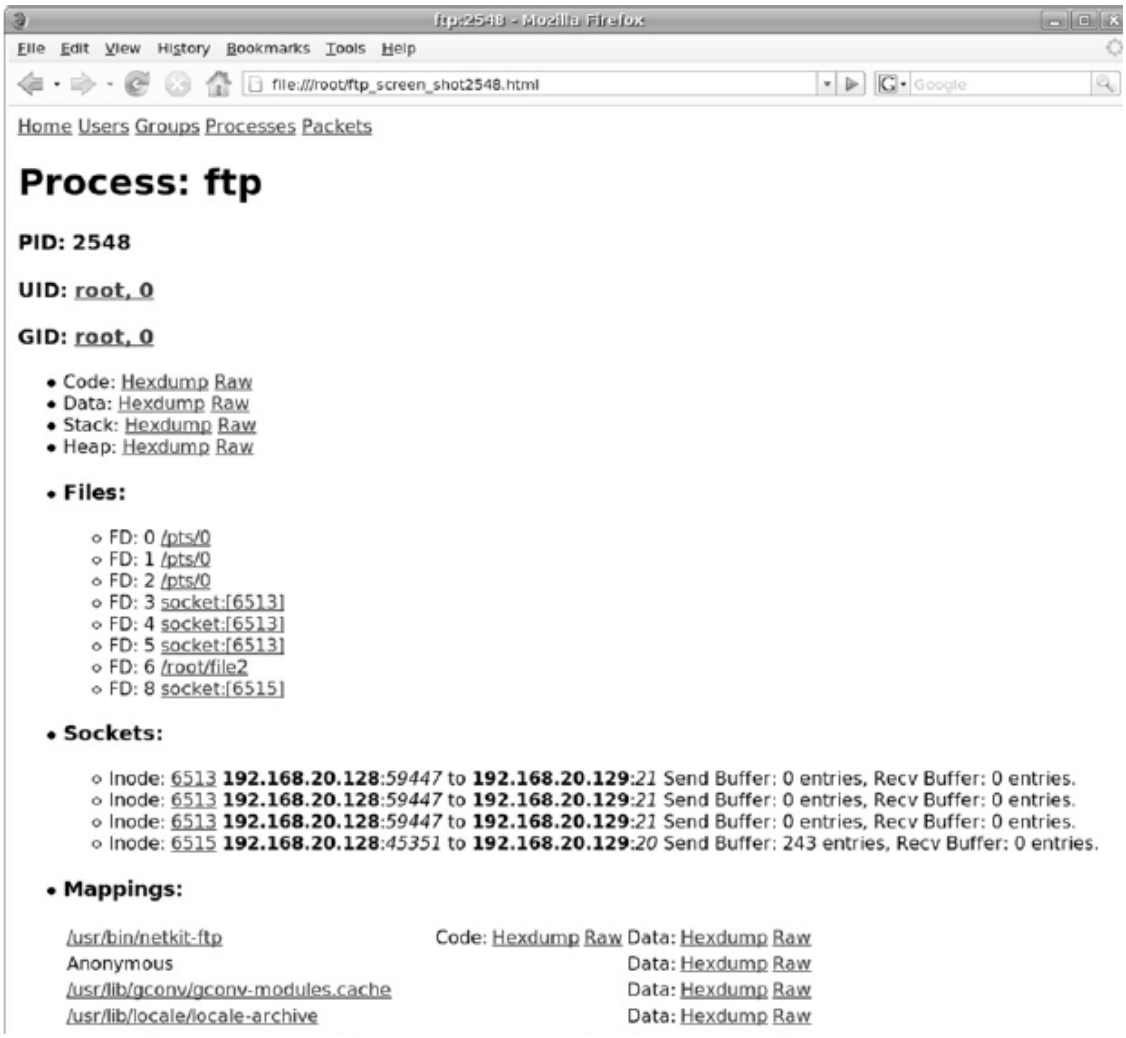

Figure 4.1: FACE framework (Case et al, 2008)

As demonstrated above, the FACE framework can be used to establish correlation between four main objects required for forensic analysis. These include the information obtained from the log files, image files, data pertaining to live system connections and any other relevant information. Therefore, by using FACE the correlation between these logical connections can be presented in a visually understandable manner and it can aid the investigator during the investigation (Case et al. 2008). This demonstrates that automation can help individuals to perform digital forensic investigations easily. Indeed, by using the FACE framework a lot of the investigator's manual work, such as linking artefacts from different sources and trying to present a picture of the crime events, can be saved, enabling investigators to perform their analysis in a more thorough and efficient manner.

4.4.6 Timeline Analysis and Automation

Timeline analysis refers to the process of reconstructing the sequence of events that culminated in the particular crime being perpetrated (Araste et al 2007). It is required to answer legal questions in courts of law to understand what has happened and to recall the scenario of the crime that is under trial (SANS, 2012). This is a complex task, mainly because the time stamp information cannot be relied upon to guarantee that the order of events is as portrayed. This could be caused by several internal or external factors, leading to a loss of reliability in such a reconstruction (Schatz et al 2006).

In comparison with other processes of the digital forensic investigation (e.g. data collection), the use of automation still has huge room for improvement. Indeed, manual methods are still used for timeline generation. The process consists of manually constructing two timelines based on cause and effect relationships, which are evident from the digital crime scenario and the available time stamps. These two sequences are then compared to each other. If there are too many anomalies between them it shows that some tampering was performed internally, or by some external intruder (Eiland 2006). It would be advantageous if future research focused more seriously on this aspect of digital forensic examination and discovered ways and means to automate this intelligent process. Such an automated process should also ensure that the timeline information generated, meets the parameters for information assurance, namely: accuracy, authentication, integrity and accountability (Hosmer, 2002).

4.4.7 Autopsy

Autopsy is one of the leading end-to-end open source digital forensics investigation platform. Thousands of law enforcement and corporate cyber investigators around world use it to find answers for cybercrimes on a daily basis. It provides a user friendly interface and its core functions are built based upon the Sleuth Kit that is a collection of command line tools, allowing investigators to analyse disk images and recover deleted files (Carrier, 2016). Autopsy is designed with several key features:

- Extensible: as it is open source software, users have the ability to add new functionalities whenever it is required.
- Frameworks: Autopsy offers standard digital forensic approaches for data processing (supporting data from different file systems), data analysis and reporting functions.
- Ease of use: it is ease of use due to its user friendly interface.

In comparison with other open source forensic tools,

By using the Autopsy, all standard procedures of a digital forensic investigation can be performed, including case management, data carving, keyword search, registry analysis, timeline analysis, and report. A screenshot of the Autopsy is presented in Figure 4.2.

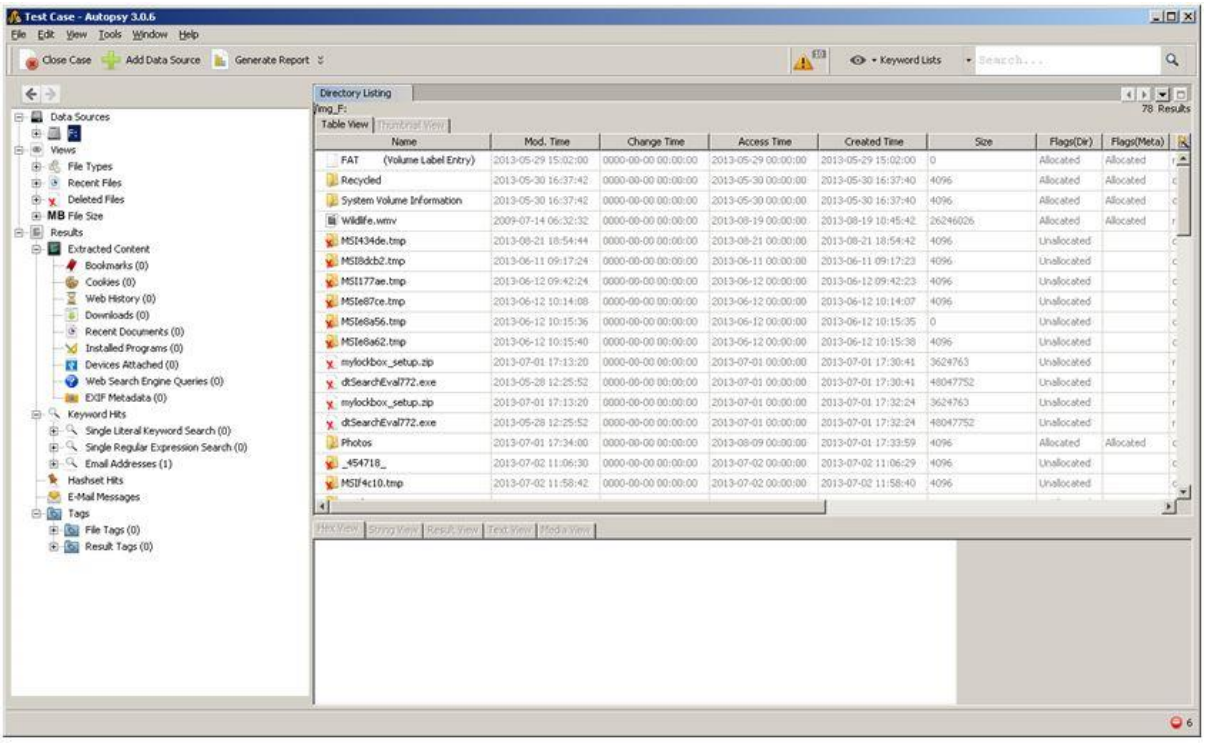

Figure 4.2: a test case of Autopsy

In comparison with other open source forensic tools, Autopsy can operate on both windows and Linux systems. Also formal support and training on Autopsy can be acquired from the Basis Technology (2016), which majority of open source are lacking with.

4.4.8 bulk_extractor

Bulk_extractor is a novel computer forensic triage tool that processes a case image and extracts useful information without needing to parse the file system or file system structures (Garfinkel, 2013). It is a powerful tool for the analysis on huge amount of data with high performance data carving, feature extractions and fast exploitation of case data. It scans the entire case data from the beginning to the end without checking the disk head and can be fully parallelized. Also the results can be easily interpreted and processed by other automated tools. An overview of the bulk_extractor is presented in Figure 4.3.

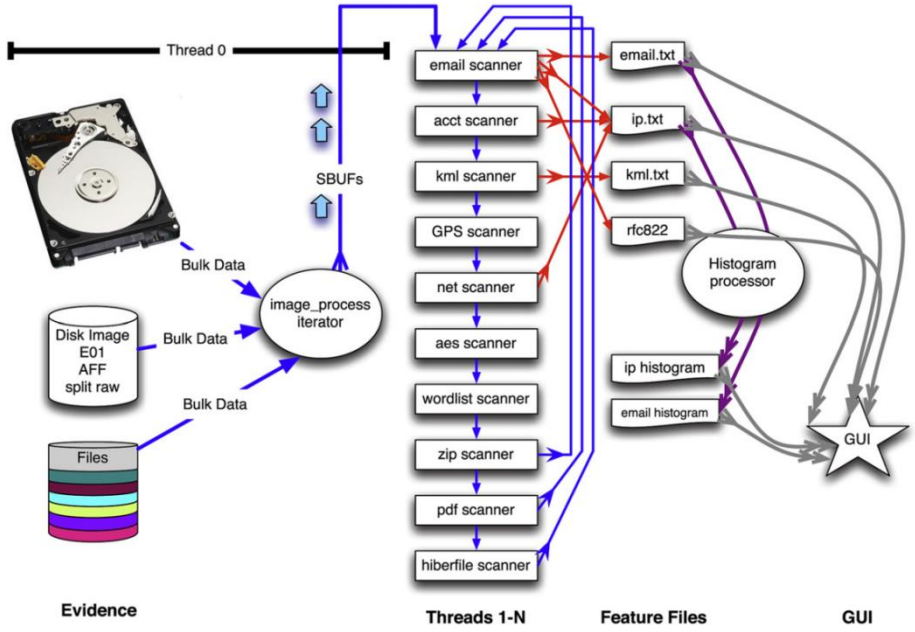

Figure 4.3: Overview of bulk_extractor architecture (Garfinkel, 2013)

In terms of digital evidence, it can extract many types from the raw image data with a high accuracy; these evidence types include email address, credit card numbers, URLs, pictures, and network information. Also it supports various case image formats, such as the raw and E01. Due to its high speed extraction speed and parallel capability, it can perform a quick triage on large amount of data within a short frame of time. For example, initial assessment on a case involving 20 laptops can be completed within 8 hours if enough hardware is provided. As a result, since its introduction, it gained a lot of popularity from law enforcement, defence and cyber-investigation applications. Moreover, the software can be used on all the main operating systems, including windows, Linux and Macintosh OS X.

4.4.9 AccessData FTK

The Forensic ToolKit (FTK) contains a suite of forensic tools (including FTK, registry viewer, and FTK imager), produced by AccessData that is one of the leading commercial companies in the fields of digital forensic and e-Discovery (AccessData, 2016). It is one of the most accepted tools by the law enforcement and the court. The investigator can use it to cover all the processes of a standard digital forensic investigation, including data acquisition, evidence examination, and report production; also it can be used for conducting different tasks, such as data carving, data decryption, registry analysis, timeline analysis, and mobile data viewing. An example of how it can be used for analysing email item is illustrated in Figure 4.4.

Figure 4.4: Case example of AccessData FTK

Although the FTK is one of the most reliable digital forensic tools on the market, it is very expensive and does require the user to take a formal training. Also, majority of the analysis is manually carried out by the investigator, which is a time consuming task.

Below is a tabular representation of the tools discussed in this chapter, listing them in terms of their main functions and categorisation. It may be seen that most of the tools listed here belong to the "commercial" category; hence there is a need to develop more open source tools as well. It can also be seen that the tools vary in their capability, and most of the tools focus on tasks that relate more to data collection and sorting, rather than any actual analysis which corresponds to the observations of Ayers (2009) as they mostly help in extracting files, hidden data, deleted files and memory dumps. While these functions are certainly important, the tool capabilities currently need to be enhanced.

Name of Tool	Function in DF	Misc Comments
Triage Concept (Rogers et al, 2006)	Quick evidence finding Find victims at risk Find the danger level of the suspect	Computer Forensics Field Process Triage Model
Case Based Reasoning (Horsman et al, 2011)	Decides best suited case on situation basis Case knowledge applied to current case Proposed case is adapted Cased added to the knowledge base for future reference	Concept which used knowledge derived from a knowledge base Based on four R's Retrieve, Reuse, Revise and Retain
X-Live (Lee et al, 2010)	Focuses on circumstances when data is huge Focuses on automation Gathers and presents data in XML format	Implements Digital Forensics XML for live data collection
Automatic Windows Log (Murphey, 2007)	Helpful in situations when direct information like time stamps etc.is not forthcoming Creates automatic log without manual intervention	Recovery Repair Validation Collation Works on Windows NT based systems e.g. XP and Windows 2003
B-Method Gladyshev and Enbacka (2007)	Useful in situations when deliberate tampering has been done with files to hide computer crime traces	Detects anomalies based on inconsistencies in data logs Based on popular systems development methodology
FACE Case et al (2008)	Presents 5 data views to users namely: Users, Group, Processes, File System and Network Capture	Uses Ramparser Linux based tool
Timeline Analysis (Araste et al, 2007)	Creates timeline with 4 parameters: Accuracy, Authentication, Integrity and Accountability	Attempts to automate timeline analysis and eliminate internal/external errors Attempts to reconstruct events that occurred during a computer crime to create a full picture
Autopsy (Carrier, 2016)	Case management, data carving, index searching, timeline analysis, reporting, mobile images	One of the best open source forensic case management tools; Also with dedicated support and training resources available
Bulk_extractor (Garfinkel, 2013)	Extracts various data types from raw images at a very high speed and accuracy, including email address, credit card numbers and URLs	A high performance triage tool gain a lot of popularity from law enforcement

Name of Tool	Function in DF	Misc Comments
AccessData FTK	Supports most of the functionalities for digital forensic investigations apart from analysis automation	One of the leading commercial digital forensic software

Table 4.1: DF Tools Comparison

4.5 Discussion

As already discussed in this section, some tools and techniques partly automate processes that help to save the time and effort of digital forensic investigators. Without this level of automation, the process of digital forensics would not withstand the immense rise in the number of computer crime incidents and the growing volumes of data that needs to be examined.

Although automation does exist in different portions of the current tools, there is a need to enhance the level of automation. Without the use of such automation, it is going to be extremely difficult for forensic examination to deal with the rising volumes of cyber forensic cases, the rising variety of devices used and the widespread use of the Internet (Hunton, 2009).

The field of digital forensics is extensive, and there are many areas to be covered. As such it is not easy to label any tool, technique or methodology as the best when compared to the rest. As it is already envisaged, most of these tools have their own sets of capabilities and limitations.

These limitations are related to various aspects, such as a lack of suitable automated tools which help to replace human intelligence, the absence of strong input from various digital forensic investigators to build a solid database which is universally available, acceptable and accessible. Many tools can only be used for some specific aspect of the digital forensic examination process, and there is a scarcity of tools which act as a one stop shop for solving various aspects of the problem.

Due to the tools' limitations, researchers have been constantly improving the techniques of digital forensics, including attempts to minimise manual effort and increase automation. Whether it is case based reasoning or case profiling, they all reflect the direction in which researchers are heading. Attempts to automate have been partially successful, mainly because there are certain areas still wherein human intellect, judgement and reasoning cannot be matched by automated processes, algorithms or computer based models.

This is important too, since the ultimate effect of the final legal outcome of the digital investigation would come to bear upon a real human being, whilst the machines, networks and technology are only a medium for the perpetration of such crimes. Hence, it is necessary for the results of such investigations to be as accurate as possible, so that no innocent person is victimised because of inaccurate findings.

The live data forensic system is a great attempt at dealing with live forensics. Although the project is still in the development phase, it is already quite extensive and comprehensive, and tries to address various requirements of a digital forensic investigator such as speed, accuracy and a degree of automation. There is certainly a great deal of improvement to be made in the system and the potential cannot be underestimated.

The triage concept has also been discussed, which is useful in situations where time is a critical factor and the quick collection and analysis of the evidence is required to have a speedy delivery of justice. The concept of triage consists of various advantages as explained earlier which helps to ensure that in sensitive cases such as paedophiles and kidnapping, valuable time is not lost in comparison with traditional investigation processes.

From a broader perspective, it may be stated that every type of tool, technique and technology has its place in the development of digital forensic investigation and no single tool or technology can be isolated as being the most useful. It is the culmination of these tools and techniques that will help in the overall advancement of the field of digital forensics.

4.6 Conclusion

One of the main features of future tools is to stress the methods of automation that are necessary to ensure that digital forensic investigators spend less time on each case and reduce

the backlog of cases. Significant effort has been put in to increase automation. However, those attempts have to be further enhanced to deal with the future needs in the digital forensic field. The process of triage is useful to ensure that speedy investigation is carried out and to save time as compared to traditional tools and techniques. On the whole, the field of digital forensic investigation has grown technologically but efforts need to be made to ensure improvements in related processes such as triage and automation are also developed.

To conclude it may be stated that though digital forensic investigation has advanced from traditional forensics, there is a lot of room for improvement, as it is still in its early stages. From the technology used for the investigation of computer crimes, to the prosecution of a victim and the legalities and legal definitions, unless they are as universal as the concept of cyberspace itself, room for error will always remain.

5 Digital Forensic Challenges

This chapter examines the current and forthcoming digital forensic challenges, including: technical, legal and resource challenges. It also includes the methodology, results and discussion of a survey study that investigates users' perceptions on how they determine and prioritise the challenges within the digital forensic domain.

5.1 Introduction

The difficulties faced by digital forensic investigators in handling computer crime cases can be categorised into several groups: technical, legal and resource challenges. Details of these challenges that need to be addressed during a digital forensic investigation are presented in the following sections. Also, a focus will be given on the new types of challenges that could hinder the digital forensic process in the very near future.

5.2 Technical Challenges

With growing incidences of computer crimes, the digital forensic tools should also be developed at the same pace (if not faster) to provide sufficient support for dealing with these attacks. At the same time, it is critical that investigators should be equipped with adequate tools and skills, allowing them to gather enough evidence for the prosecution of the perpetrators. Karyda and Mitrou (2007) discussed the technical and legal challenges being faced during digital forensic investigations. The diversity and heterogeneity of the infrastructure and physical barriers involved prohibit investigators from accessing the sources of evidence.

Most models work on a presumption that an attack has taken place in order to apply certain procedures, with the objective of discovering and collecting relevant evidences. To begin an investigation, a deep understanding of the characteristics of the attack is required. Hence, the case can be dealt with accordingly. Important challenges that forensic investigators and models need to take into consideration include the growing size of data storage, the prevalence of embedded flash storage, the need to analyse multiple devices, the use of encryption and cloud computing (Garfinkel, 2010; Moore, 2006). Choosing between the more important and relevant information is a further challenge when dealing with a large

amount of data. Also, large networks over multiple systems are another difficult proposition for digital forensic investigators to overcome. Multiple systems (using Network Address Translation) sharing the same IP address further poses a challenge as it is difficult to relate the traffic to a specific host (Cohen, 2009). Another difficulty that the Internet poses is in terms of conducting date and timeline analyses on collected data. Mohay (2005) discussed both monitoring the Internet and large volumes of data as challenges to be dealt with by digital forensics investigators.

Cybercriminals utilise a wide range of techniques to avoid being traced and captured by authorities. They create various obstacles, with the objective of removing the evidence or to cast a shadow of doubt on the evidence collected. In E-crime Watch Survey (2006), changing file extensions, utilising swap space and disk wiping software, physically destroying media, techniques facilitating anonymity, cryptography and steganography were recognised as common activities.

The current practice in digital forensic principles, tools and practices assume that the storage media is under complete control of the investigator (Grispos et al, 2011). This ignores the challenge posed by online storage of data that can be easily exploited by perpetrators.

Information technologies in today's era have a dynamic nature. For instance, according to Moor's Law, information technology becomes obsolete every 18 months, resulting in the unstable and unpredictable environment for continuance of the same infrastructure (Moor's Law, nd). This means that the advancements in technology are at a fast pace which makes the underlying infrastructure obsolete for utilising the full potential of advancements. Digital devices, notebooks, iPods, mobile phones, cameras have developed very quickly. According to Mohay (2005), it is a challenge to keep up to pace with new devices when developing appropriate tools. This further affects the digital evidence to be acquired. Also Bogen and Dampier (2005) stated that a single tool which is capable of meeting the entire set of needs of an investigation does not exist.

A further challenge for digital forensics is forensic readiness – which is the ability of a system to capture and use evidence in an effective manner (Endicott-Popovsky and Frincke, 2006; Yasinsac and Manzano, 2001). Forensic readiness is a term used to describe the extent to which computer systems and networks record data and activities in a manner that ensures a

sufficient record for subsequent purposes is maintained. It is important that the records can be accepted as authentic within digital forensic investigations.

An organisation is in a much better position to handle a digital crime incident if it is in a state of forensic readiness (Casey, 2005). This refers to various factors such as properly trained handlers and so forth. It is possible that after an incident has taken place, the first respondent might unknowingly damage the evidence or carry out some activities, which makes it difficult for the forensic team to gain clues regarding the incident. As a result, evidence needs to be handled with care, and even a trivial item needs to be kept on record. In the absence of such close scrutiny it is possible for evidence to be destroyed or never found in the right place at the right time. Being forensically ready is also important because in most cases it is the same set of personnel who are both the incident handlers and forensic investigators.

Cloud Computing has drastically changed how the creation, delivery and management of information technology services are conducted (Ruan et al, 2011). The main categories of the types of Cloud services are described as follows (Grispos et al, 2011):

- In the case of a private Cloud, it is most likely that a single organisation owns and controls the Cloud infrastructure, and everything is located in the same geographical location.
- Community Clouds are relatively more dispersed, and tend to collectively use the resources of Cloud Computing.
- Public Clouds are meant to be used for various users and hence contain data from a variety of users.
- Lastly, a hybrid Cloud is a mixture of one or more of the above models.

Hence it can be seen that the technology of Cloud Computing, acts as an excellent technical platform by providing distributed services at economical costs, allowing people and organisations of all scales and levels to reap the benefits of software, packages and services which would otherwise be beyond their reach. However, there are several challenges associated with Cloud Computing from the perspective of digital forensics due to its vastly distributed nature. For instance, it is not possible to demarcate precisely the area of operation of the “Cloud” especially in case of public Clouds. Within Cloud environments, security and confidentiality are major concerns along with issues such as encryption, proliferation of endpoints, multi jurisdiction, and loss of control over data (Curtis et al., 2010). Reports of

Botnet attacks on the Cloud infrastructure of Amazon and the recent hacking of Gmail illustrate that the Cloud environment is already a target for malicious intentions (Kirwan, 2013).

Furthermore, since the Cloud is embraced globally, the number of computer crimes committed in the Clouds will rise too.

5.3 Legal Challenges

According to Kenneally (2002) digital evidence constitutes a unique legal challenge. Reliable evidence that can stand the rigorous requirements of admissibility is critically important to computer crime investigations. Admissibility of the evidence is dependent on its relative weight and judicial value. Legal systems are based on precedents and this raises the need to introduce cohesion and consistency in the adopted digital forensic methodologies.

Allen (2005) and Wilson (2008) raised the issue of jurisdiction being a challenge for digital forensics. Features like portability and connectivity raise the question of jurisdictions and the difference between legal systems in different jurisdictions complicates the matter further. For a piece of evidence to be considered admissible, it must fulfil the evidentiary mandates of the court in which it is presented. Lack of effective cyber laws is another major obstacle in investigations. For example, no legal action could be taken against the author of the 'ILOVEYOU' virus in the year 2000, as the suspect was located in the Philippines and the country had no legislation with respect to such crime at that time (Karyda and Mitrou, 2007). Another important challenge that is critical for digital forensics is the issue of privacy. For instance, seizure of equipment and release of information are often disputed in the Court on these grounds.

In the case of Daubert vs. Merell, the US Supreme Court provided specific criteria for the lower courts to rule on the admissibility of scientific evidence (Rogers, 2003):

- Whether the theory or technique has been tested on reliable grounds;
- Whether the theory or technique has been reviewed and published;
- Has the theory or technique been tested and analysed for potential errors;
- Whether the theory or technique has been accepted by the scientific community.

Hence the court stresses the reliability of the evidence; and if the evidence fulfils this criterion, it would still be accepted in a court of law. However, this does mean that anything is permissible and benchmarks will certainly exist.

A further major legal hurdle in the USA is the use of evidence in courts. The Fourth Amendment (Adams, 2008) states:

"The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized".

This acts as a significant barrier in the search and seizure operation because the computer hardware and other related accessories which are taken for digital forensic analysis normally contain a huge amount of personal and private information, apart from the expected data that could lead to the perpetrator.

Search warrant issues for the forensic investigator are too restrictive at times and may even specify the types of files they are supposed to search for on the suspect's computer. Such facts make the task of a digital forensic examiner much more laborious than in the case where a generic search warrant is issued (Adams, 2008). It also means that it is the choice of the examiner as to whether to exhibit strict integrity and professionalism by not obtaining anything which is unrelated to the warrant. Hence, considering various legislative aspects is a key requirement in the next generation of digital forensic tools, allowing the investigator's boundaries to be pre-defined.

It is understandable that the state of cyber laws is not equal everywhere in the world. For example, taking the case of the Sultanate of Oman where the decree implementing cyber laws was passed as recently as February 2011 (ITA, 2011) and the document describes various aspects relating to cyber laws. Although the document tries to incorporate various facets of cyber laws, it is still not as comprehensive as places where a system of well-established cyber laws exists, such as the United Kingdom. The cyber laws in the UK are much more refined and cover aspects relating to various dimensions like digital signatures, e-commerce, intellectual property rights, and Cloud Computing; in addition it is a signatory to various

international conventions such as the convention on computer crimes and EU data protection directive (BSA, 2012).

Cloud Computing also raises a different set of challenges for digital forensic examiners due to its dispersed nature of data; as a result it is difficult to pinpoint the exact legal framework that would be used when incidents occur within a cloud environment (Ruan et al, 2012). Also as the controlling personnel may be different from the company or individuals whose data is held in such Cloud environments, it becomes more difficult from the legal perspective to retain the evidence in an untouched and unaltered manner, suggesting that the evidence could not be accepted as solid proof in a court of law.

According to Alazab (2013) there are five main objectives of digital forensics which are listed as follows:

- Any undesired events should be detected;
- The impact of these events on the system should be analysed;
- Requisite evidence should be gathered from the legal perspective;
- Ways to prevent such future mishaps should be devised;
- The pattern and underlying reasons for the event should be analysed and understood in order to prevent any future happening of such events.

This view includes evidence that may not be admissible in a court of law. It is possible to collect evidence from events like theft or destruction of intellectual property, commission of fraud, or any such criminal activity relating to digital devices. Digital evidence can also be used for establishing a link between the crime and the victim (Perumal, 2009).

Carrier (2002) also discussed a number of scenarios that involve the use of open source toolkits for carrying out digital forensic examinations. This trend has become more prevalent after the involvement of commercial organisations in the forensic investigation process, since earlier it was mainly a governmental matter. Therefore, it is necessary to ensure that the quality of digital tools and the evidence which they produce are technically sound and without any flaws, which consequently makes it difficult for a court of law to accept such evidence.

Also, an open public debate should be organised in tune with the open source tradition, so that there is a consensus on the type of tools being used in digital forensic analysis, the methods used by these tools and the acceptability of their results.

Hence it can be seen that though the field of digital forensics pertains to procuring and presenting evidence considered admissible in a court of law, it is not so easy in actual practice. There are many hurdles and challenges which a digital forensic investigator has to face.

5.4 Resource Challenges

First and foremost, the challenge in terms of resources to be dealt with includes that a clear definition of digital forensics has not been provided in creating differences in the objectives, standard properties and the methodology to be followed and adopted; making it difficult to make a formal comparison of systems (Carrier, 2006). There is no recognised methodology that enforces forensic capability (Taylor et al, 2007). It is important that as the field grows, resonating advances are made in terms of capabilities specified, implemented and verified. To reach an agreement on definite and formal methods is important so standardised properties are defined.

Certification of people, tools development, ongoing laboratory experiments and researches are important aspects to fulfil the future needs of digital forensics (Garfinkel, 2010). The need to develop special courses with appropriate focus in terms of university degrees and industry courses is imperative (Barbara, 2005). Training of staff involved in forensic investigations is important for seeing successful digital forensics. It is important that the investigators are trained in a manner that allows a complete examination of the scene while preserving volatile evidence, be able to work with a variety of digital devices and integrate the evidence with the investigation.

A further challenge in this arena is the non-uniform availability of resources available to investigate and fight computer crimes especially in developing countries.

Gercke (2011) claimed that cyber criminals could take all these factors into account while launching many major cyber attacks. Cyber criminals are well aware of the fact that only

limited resources are available to fight computer crimes in the developing countries of the world. Even if one got caught in such a place, it is relatively easy to escape from legal bindings, as compared to a same situation happening in a developed country, where the implementation of such laws are mature and fairly strict.

As demonstrated, the use of computer technology and the Internet expanded significantly and consequently the number of computer crimes has also accordingly increased, posing new challenges to the authorities and organisations combating them.

5.5 Future Challenges

It is widely understood that the past and current trends and studies can be used to find the generic direction in which research is heading. The trends of computer crime show a consistent rise over time; indeed, one study suggests that the number of computer crimes increased 100% in the past three years (CFS, 2012). If this situation continues, the world would certainly see a big rise in computer crimes especially related to those on Internet enabled mobile devices (Norton, 2012).

In addition to the challenges presented by Cloud Computing discussed earlier, a survey by Ernst & Young (2011) discovered that nearly two thirds of the companies they surveyed considered switching to or utilising Cloud Computing services in future. This would certainly lead to risks associated with data integrity and could complicate legal perspectives involved in such cases. There would be an increasing need to implement multi-dimensional service level agreements between the Cloud service provider and the Cloud consumer which ensures that the relevant legislations are addressed and implemented, involving different jurisdictions of the Cloud service provider, the client and the place where the actual data resides (Ruan et al., 2012). For instance, when a US company's branch located in the UK intent to use a German Cloud provider's service, the service level agreements need to be created in a way following various rules and regulations, including the US company's policy on data, EU laws regarding how the data is processed and stored by the service provider and UK's regulations on computer security. In fact, this has been seen as one of the most important future challenges acknowledged by RSA (2012), to ensure international cooperation amongst the different law enforcement agencies of different nations, and the need to improve information sharing which helps to crackdown on criminals across geographical and political boundaries.

Some of the recent surveys discussed previously, conducted by Norton (2011), McAfee (2012), RSA (2012), Ernst & Young (2011), BAE (2011), and Ponemon Insitute (2012) point towards the fact that the computer crimes would certainly pose increasing challenges to digital forensics in the near future, including

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

According to a report of RSA (2012) cybercriminals are becoming better equipped with sophisticated technology. For example, software packages such as Zeus are emerging as hugely popular tools on the Internet black market; they have advanced algorithms to break security and can be used in financial crimes and frauds.

Another future challenge pointed out in a study by PwC (2011) is that companies would need to consider in-house security as well as external threats as nearly 20% of internal employees used their skills and resources to cause insider damage. Also, Ernst & Young (2011) suggested that organisations have taken a futuristic view and see risk management as an investment in the future rather than a preventive maintenance act in order to combat all types of threats including computer crimes.

The study of challenges in the field of digital forensics is the first step towards finding the most important areas that need to be addressed within the research context. However, apart from studying the previously available literature, it is also important to review the current state of the art in the digital forensics domain. Since the area of digital forensics is technical in nature a relevant survey would be an appropriate approach to seek opinions from those who are constantly involved in the field through practice, training or research. Therefore, a survey that can attain views from different aspects is necessary. To this end, a survey was conducted in particular to pinpoint the gap in knowledge and to decide on the most

appropriate approach to contribute in solving the problem within the domain. Details of the survey (including the purpose and methodology) are fully described in the following sections.

5.6 Survey into Digital Forensic Challenges

An overview of the difficulties in the digital forensics domain in the previous sections showed that they pertain to different areas such as technical, legal and resource challenges. It was also seen that these challenges are predicted to multiply in the future given the rapid development in technology. Indeed, Internet enabled devices are increasing in both functionality and popularity; in contrast, decreasing in prices makes them affordable for most people. Such a fact has indicated the need for researching in the various challenges from different aspects, in order to propose a correct means to counter and manage computer crimes. Whilst prior research has focused upon the challenges which exist within the domain, a lack of literature exists in validating these challenges, and in understanding the differing stakeholder perspectives; prioritising these challenges will focus the research effort.

Thus, this survey investigated the challenges from a literature point of view and attempted to associate the findings with feedback from the relevant stakeholders. Also, the survey intended to seek feedback from participants pertaining to the future challenges of digital forensics. This has led to canvassing practitioners and researchers who have the requisite skills, experience and expertise to put forth an intelligent opinion on the matter.

5.7 The Survey Methodology

The purpose of the survey was to better understand the collective and individual perspectives of people related to the digital forensics field, including academic researchers, law enforcement practitioners and non-law enforcement practitioners (e.g. individuals working in organisations in a digital forensic capacity). The survey was constructed with quantitative analysis in mind. As such, the survey was structured in the form of multiple choice questions which aimed to extract reflective feedbacks and opinions from respondents within a relatively short period of time. It was envisaged that this approach would maximise the number of completed surveys.

Formally, the aim of the survey was to establish and prioritise the key future challenges that are posed within digital forensics from various perspectives of practitioners and researchers. In order to achieve this, the following objectives were defined:

- For establishing a baseline, understand the general forensic background of participants – in order to appreciate the relevancy of responses;
- Understand what participants feel the role of digital forensics is and to what extent it is currently able to meet that requirement;
- Understand what (if any) impact changes in future technology will have upon digital forensic investigations (from both a law enforcement and organisational perspective);
- Examine what effect the evolving nature of digital forensic capability and knowledge will have upon practice – the evolution of anti-forensic techniques and technologies that are security aware and reduce the forensic opportunity (e.g. the operating systems that forensically wipe files upon deletion).

The format of the survey was kept simple yet effective, so that there was an optimum balance between the time spent on the survey, the efforts required to complete it, while giving enough room for the questions to let the participants express their views in the most succinct and reasonable manner possible. To achieve this, a number of trial runs of the survey were undertaken by local researchers in the domain acting as participants. From the feedback obtained, the survey was further refined and optimised. The pilot phase was designed to seek feedbacks regarding the survey from the researchers within the domain. This feedback was then considered in order to optimise the survey feedback from the 'real' participants.

The survey was designed in a structured format, starting with a demographic section and moving on to the role of digital forensics, future challenges and finally the legal aspects of digital forensics. The demographic section provides a brief overview of the respondents in terms of relevant qualifications and experience in the field of digital forensics. This gives an idea of the gravity and experience of the opinions being given. The next section attempted to gauge the overall perspective of the current attitudes of digital forensics by asking questions relating to the importance of digital forensics from different perspectives and also the limitations of the current tools and techniques. The section pertaining to future technologies was designed to ascertain how professionals in different fields feel about future trends, and

which areas require improvements. This section also focused on confirming and prioritising these challenges. Finally, the section relating to the legislative aspects was designed to investigate what professionals feel about the legal perspective of digital forensics.

The survey was published online, and the researcher shortlisted potential participants chosen from the different areas identified above, who were emailed directly. In total, 128 invitations were sent out to the potential participants. As for the researcher's category, they were approached through the association and links the researcher had within the Centre for Security, Communication and Network Research (CSCAN) and the Computer crime Forensic Specialist Group (BCS), while the practitioners were contacted through Oman National CERT. A total of forty two participants have successfully responded and completed the survey. Given the highly targeted nature of the survey, the response level was outstanding and provided a solid basis for the survey analysis (Appendix C).

5.8 Survey Results

The raw data obtained during the survey was sorted and analysed in order to explore the demographics. The first three figures were plotted with different demographic factors. As can be seen from Figure 5.1, the vast majority of the respondents had postgraduate qualifications. In general, this shows that the respondents were very well educated, and provided a good basis for respondents to understand the domain and its associated challenges.

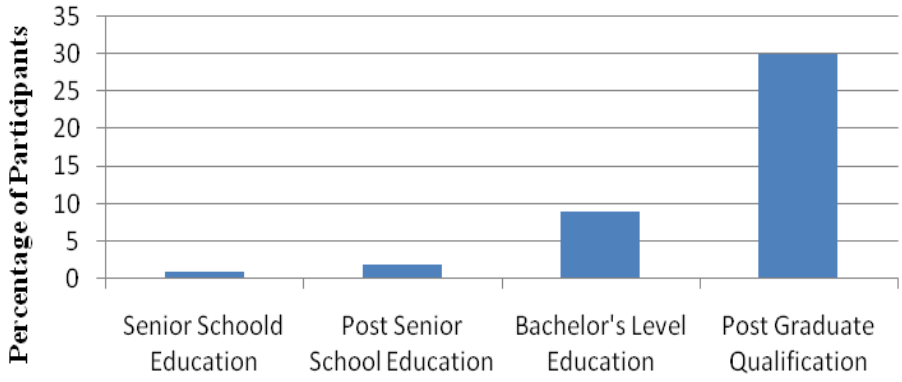

Figure 5.1: Education of Respondents

Similarly, Figure 5.2 gives the percentage of the participants in terms of their background, whether they are from the academic community, law enforcement or organisational security. As can be clearly seen from the graph, 45% of the respondents belonged to the academic community; also a significant proportion of them (i.e. 47%) were directly or indirectly

involved in digital forensics, either by being law enforcers or members of organisational security teams; 8% of the members were drawn from other areas. This certainly lends an edge of credibility to the sources used for the survey, since the responses are drawn from a pool of individuals who have sound knowledge of the subject matter through their education and work experience.

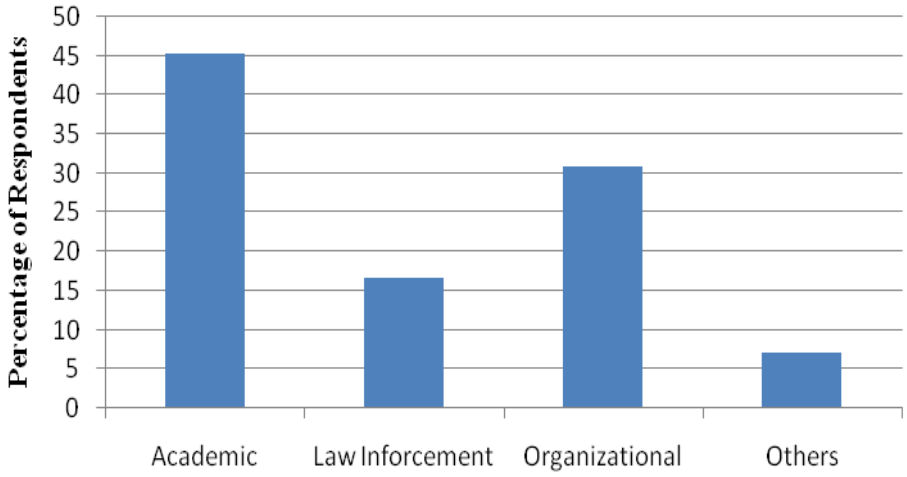

Figure 5.2: Category of Respondents

Figure 5.2 illustrates the number of years the respondents have spent working in the digital forensic field. It can be clearly seen from the bar graph that except for 8 respondents who had less than one year of experience, the remaining had been working in the field for some time; also 55% of respondents had 3+ years of experience, arguably providing a significant exposure to the domain.

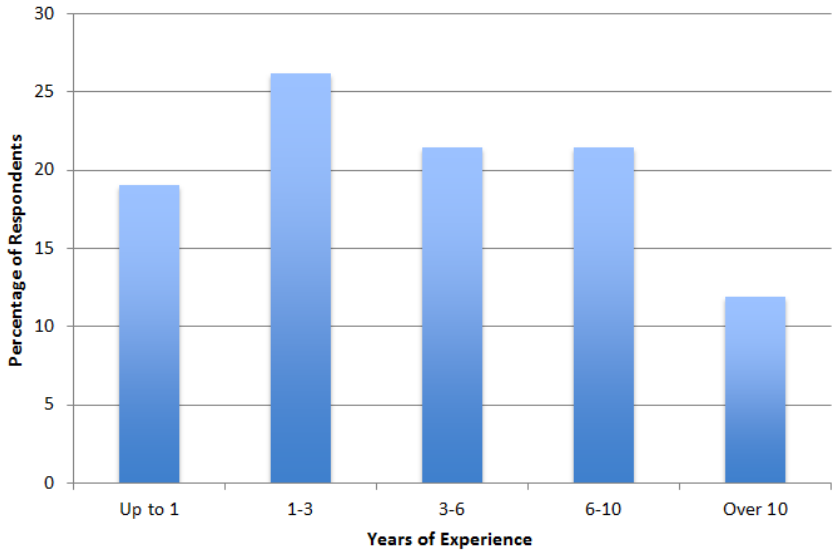

Figure 5.3: Experience of Respondents

Overall, whilst the number of total respondents (despite the high response rate) is relatively low, the demographics show those who did respond had an appropriate level of knowledge, expertise and experience to answer the questions. Moreover, having successfully obtained responses from all three stakeholder groups, this provided an invaluable insight into their respective priorities.

Figure 5.4 shows the results of how the respondents perceived the risks in terms of limitations of the current digital forensic tools, which has been broken down in terms of the category of respondents.

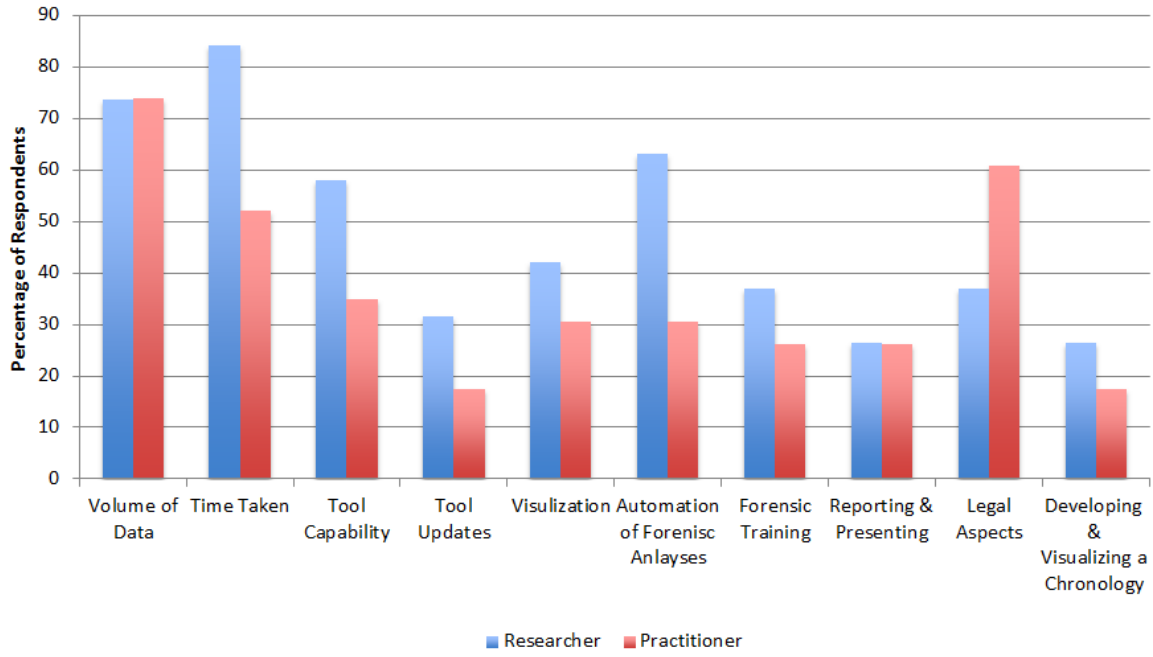

Figure 5.4: Limitations of Current digital forensic Tools

As can be seen from the graph, both the researchers and practitioners agree in equal proportion, which is about 74%, that high data volumes is one of the main limitations that current digital forensic tools are facing. Obviously this results from the increasing use of digital devices including computers in all aspects of human life which in turn gives rise to huge amounts of data. All this data needs to be managed in terms of proper storage, processing and so forth, which in turn gives rise to difficult situations which give rise to mismanagement and errors that could lead to the propagation of computer crimes.

From the researcher's perspective, the biggest challenge is the time taken to complete the digital investigations; indeed, more than 84% of the respondents strongly agreed with this. In comparison, 52% of the practitioners agreed on this point.

Automation and tool capabilities come next on the list of limitations from the researcher's point of view, with 63% and 59% of the researchers respectively considered them limiting; while fewer than 35% of practitioners agreed on these two challenges. Nearly 60% of the practitioners thought that legal hurdles are an important issue, though less than 40% of practitioners agreed on the same issue. This obviously shows that respondents with various backgrounds in digital forensics have different views about the proposed challenges. For example, the practitioners have to deal with courts, law officers and trials on a regular basis rather than academicians who are more concerned with theoretical aspects.

Regarding other factors like reporting and presentation, visualising technologies, and forensic training, less than 40% of respondents considered them as challenges for digital forensics in the future, this is understandable as the majority of them are related to the functionality.

In terms of issues related to the automation, it clearly illustrates that the volume of data, time taken and tool capability are the top three areas requires further attentions. Actually, these three areas are closely related. Due to the increasing data volumes and the lack of tool capability to deal it, more time taken to analysis is obviously required; the time factor is also affected by the large number of computer crimes exists today. In order to solve these three issues with one solution, tool automation is one of the potential candidates. As the examination and analysis process is automated, more work can be carried out within a given time; hence, the time taken and large volume of data issues can be solved. As a result, automation for the examination and analyse phase of the digital forensic investigation is a critical area that should be focused upon from the viewpoints of research and industry.

Figure 5.5 shows the views of the participants regarding the fitness of tools that are currently available for computer, network, mobile and embedded forensics on a 1-5 scale. The value 1 represents that the current tools available for the particular forensic category are hardly useful or fit for purpose; whereas the value 5 represents that the tools are totally fit for purpose and hardly require any improvements.

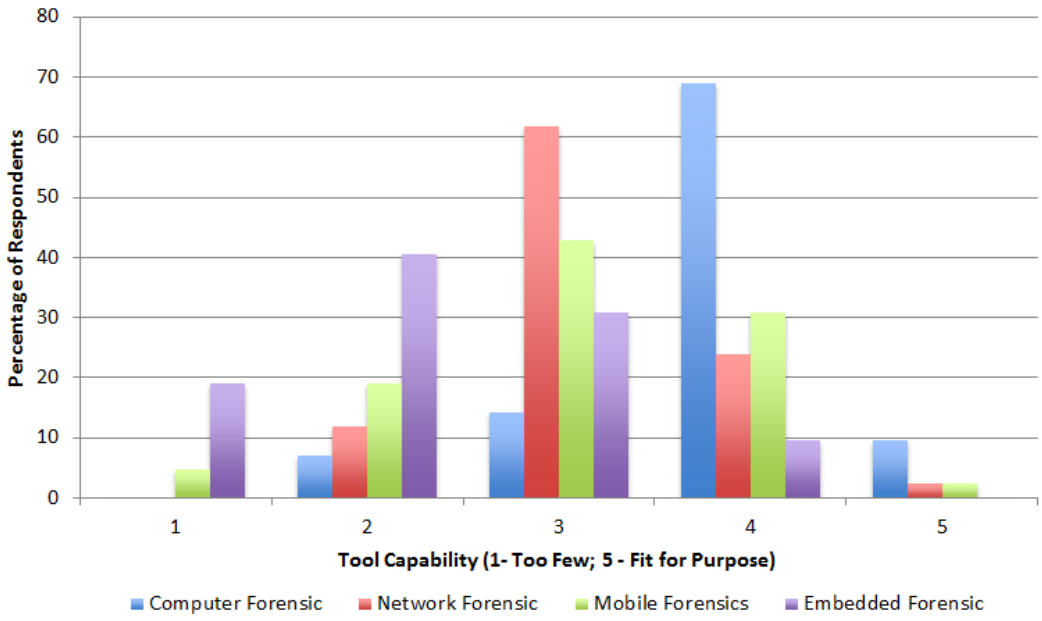

Figure 5.5: Fitness of Current DF Tools on 1-5 Scale

As can be seen from the response, there is an acute shortage in terms of the level of fitness of tools, except those for computer forensics where a majority of the respondents believed that the tools are somewhat fit for purpose with almost 70% them giving a rating of 4. This is expected as according to the literature the tools available and focus within digital forensics have traditionally been on computer forensics. In comparison, there is a need to develop new tools or improvise upon the capability of existing tools in the network, mobile and embedded forensics areas. In essence, it shows that very few individuals considered that the current tools were fully capable and fit for the purposes of digital forensic analysis, indicating that opportunities for development certainly exist across all four sections of digital forensic sub-domains.

Figure 5.6 demonstrates participants' concerns over various factors of digital forensics, including Cloud Computing, malicious software and anti-forensics. Respondents were asked to rank these factors on a scale of 1 to 5, in which value 1 meant that the factor was of their lowest concern, while the value 5 meant that the candidates had that factor as highest on their list of concerns.

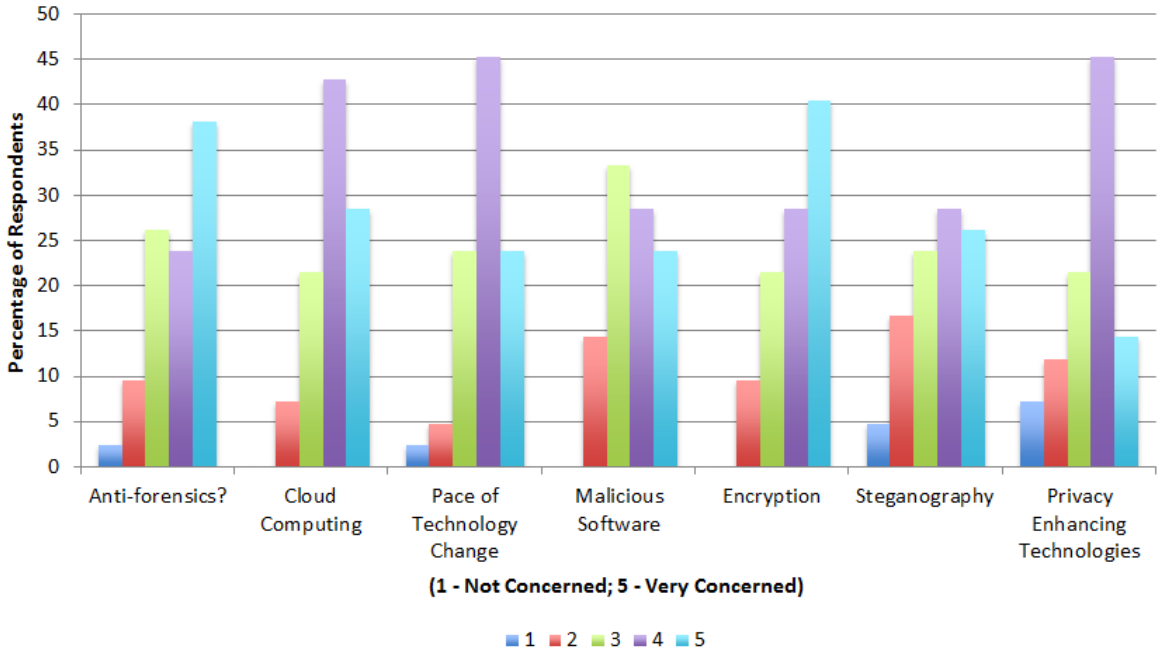

Figure 5.6: Participants' concerns

In terms of respondents being very concerned, the use of encryption, anti-forensics and Cloud Computing are the most concerned factors regarding digital forensics, with 40%, 38% and 29% of the participants respectively. Also the respondents show some level of concern regarding the use of privacy enhancing technologies and the pace of technology change. In comparison, they were less concerned about the presence of malicious software within the domain of digital forensics.

Also, Table 5.1 presents a breakdown of top concerns of both researchers and practitioners and the overall perspective of both groups combined in terms of the concerns with regards to cloud computing, anti-forensics and encryption. It must be kept in mind that encryption is also equated to the pace of technology by the researchers while they do not place much emphasis on Anti-Forensics which is on the top of the list by practitioners. It may be seen that the researchers place most emphasis on cloud computing as the main concern in digital forensics while in the case of practitioners as well as the overall group, it is the anti-forensics factor which is perceived as the greatest concern.

	Priority		
	1	***2***	***3***
Researchers	Cloud Computing	Encryption = Pace of Technology	
Practitioners	Anti-Forensics	Encryption	Cloud Computing
Overall	**Anti-Forensics**	**Cloud Computing**	**Encryption**

Table 5.1: Top Concerns of Various Factors

Figure 5.7 presents respondents' opinions regarding the importance of several factors including profiling, privacy impact and legislation for digital forensics. The participants ranked all these factors on a scale from least important to extremely important. A number value of 1 to 5 was given to each of these answers: with 1 for least important and 5 for most important. It must be noted that unlike most other graphs in this chapter, this figure on the y-axis here does not represent a percentage but a relative weighting, calculated in the above mentioned formula to show the relative significance of the responses to each other.

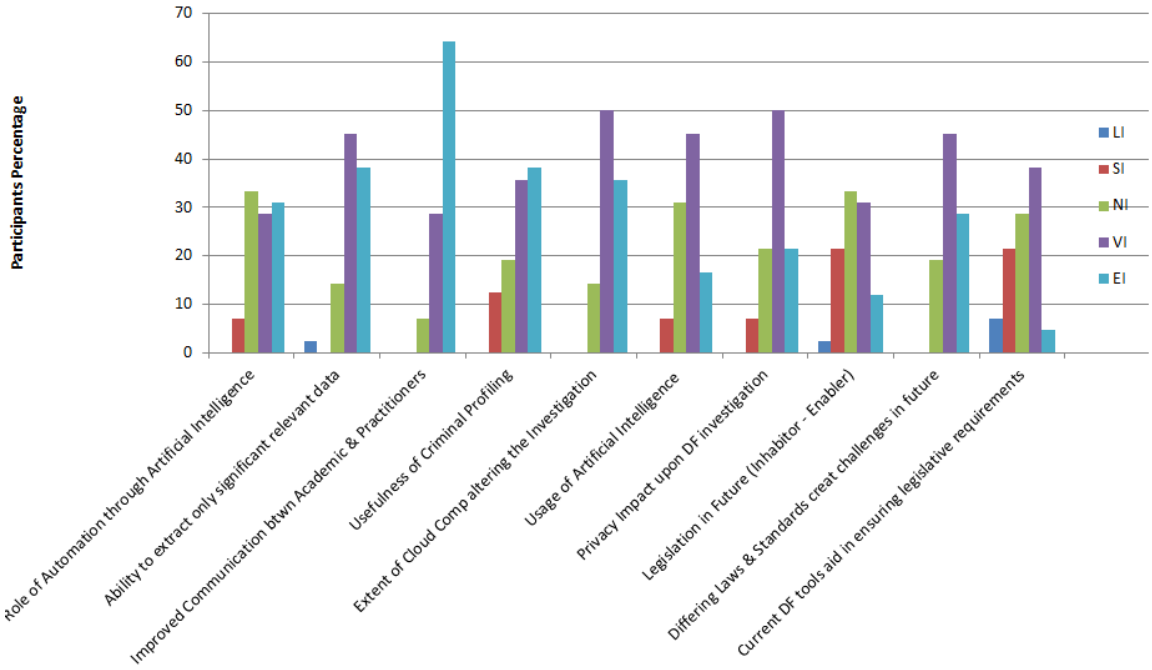

Figure 5.7: Importance of Factors in Digital Forensics

Principally, this prioritisation seeks to quantify the importance the respondents attached to that factor. It is noteworthy that the factor which carries the most weight is the importance of communications between academicians and practitioners, indicating that a communication gap exists between people who research and the ones who actually do the task practically within the digital forensic domain, and they need to understand each other clearly.

Also, the extent to which Cloud Computing altered the investigations clearly worried the respondents, as well as the ability to extract relevant information from the heaps of data that

are present. Legal scenarios were the next disturbing in order of importance, since respondents perceived the difference in such laws to be a deterrent to investigations. Criminal profiling occupies a somewhat middle position in the order of importance followed by privacy and automation impacts. Artificial intelligence also showed the importance of digital forensic investigations; as a result, there is a need to have thorough regulatory frameworks which enable the aforementioned to follow digital forensic requirements.

Figure 5.8 shows the future trends for four digital forensic sub-domains namely mobile, network, embedded and computer forensics. The results clearly indicate that both researchers and practitioners agreed that mobile forensics will certainly be an important factor in the future. This is understandable since the number of mobile devices is continuously on the rise with more multi-functional features.

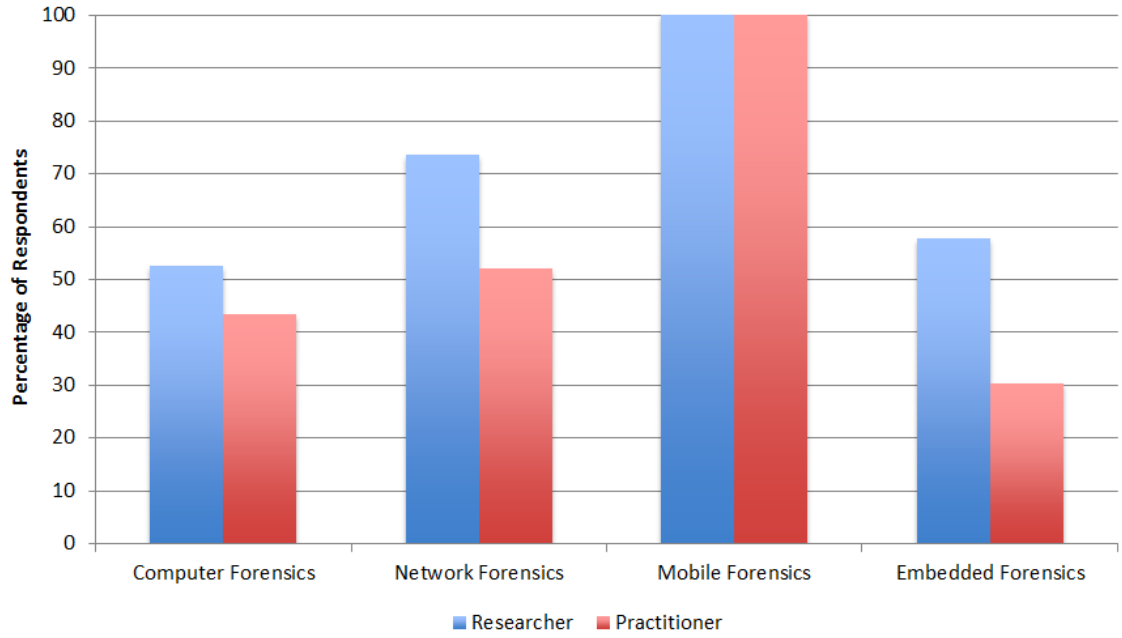

Figure 5.8: Future Trends in Digital Forensics

A proportion of 72% of the researchers and 52% of practitioners feel that network forensics will witness a rise in the future. While opinions on the future trends for embedded forensics and computer forensics are somewhat divided, although the researchers were slightly in favour of the trend than the practitioners. The results seem to be reasonable in the light of the fact that mobiles and networks are rising at an immense rate so they would certainly act as a challenge in the future. In comparison computers have been around for decades and they may have reached a point where their impact will not increase exponentially even if they have not

reached a saturation stage yet. Also, from the practitioner's perspective, embedded forensics will be less challenging since it does not have a direct interaction with the end user device as embedded devices are mostly hidden at a deeper level.

Figure 5.9 demonstrates the representation of the participants' views on legal concerns regarding digital forensics. This was based on the three main questions asked, relating to the different laws posing a challenge to digital forensic investigations, whether they acted as Inhibitors or Enablers and whether current laws aided legislative requirements.

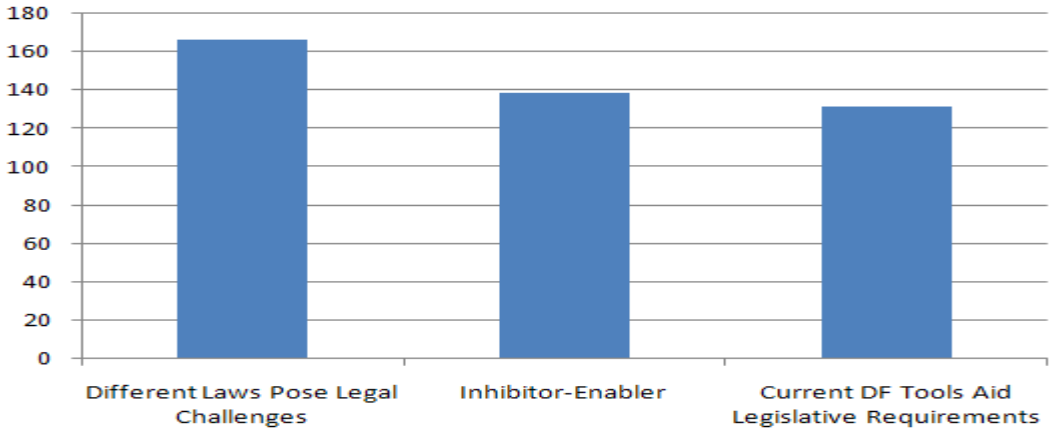

Figure 5.9: Legal Concerns

The candidates were given a choice to make their responses in terms of a 1-5 ranking and later on these responses were converted into relative weighting. In terms of percentages, 29% of respondents strongly agreed with the statement that different laws do pose a challenge to digital forensic investigations; while 45% were also in general agreement about this. In contrast only 7% of participants disagreed with this fact, showing the general perception about the regulations is not highly positive. Notably, 43% of the respondents believed that existing laws are an enabler for digital forensic investigations. Fewer than 5% of candidates strongly agree that current digital forensic tools are fully fit to aid legislative requirements. This demonstrates that as far as legislation is concerned, this needs to be improved and modified along with digital forensic tools which need to be more in harmony with one another and offer additional help enabling digital forensic investigators to bring criminals to justice more easily.

5.9 Digital Forensic Challenges and Survey Discussion

The digital forensic domain has faced a plethora of challenges relating to technical, legal and resources issues. The amount of data and the number of multi functioning devices used are

rapidly increasing. This fact poses a number of challenges to the digital forensic field and makes the digital investigators extremely busy trying to cope with the constant increase of the pending cases waiting to be analysed. Moreover, the rising use of cloud technology has introduced another challenge in which the data is not stored in the suspect computer but in a third party digital space. For example, the data stored in a Dropbox account or hosted in any similar service is not easily accessible by the investigator; as a consequence, potential evidence could be missed. The legal issues also weigh heavily against the adoption of a universal set of standards to deal with computer crimes since the traditional laws are mostly related to physical objects and boundaries and vary from country to country, whereas the Internet is literally borderless. Combined with the above challenges is the issue of resource scarcity where sufficient people with the relevant digital forensics skills are not available to examine and investigate the cases, and so computer crimes are experiencing a significant gap between the number of pending and resolved cases. As well as reviewing the challenges faced by previous scientists, the researcher also carried out a survey involving stakeholders in the domain of digital forensics to discover their views.

The survey showed that the majority of the experts believed that digital forensics would increase in its scope in the future; and although all types of forensic examinations are perceived to be of increasing importance, mobile forensics occupy a prominent role, followed by network, computer and embedded forensics.

One of the most important issues which seemed to be of concern to nearly all participants was the gap in understanding between the academicians who carry out research in the area of digital forensics and the professionals who use the tools, technology and techniques developed by the researchers in actually solving digital forensic cases at ground level. The two groups need to understand the challenges and viewpoints of the other party in order for a better evolution and implementation of the tools and forensic science as a whole to take place from the practical perspective. Since mobile devices are on the rise, the survey pinpoints that all participants, without exception, are of the view that the area of mobile forensics is set to see a growth in the future and will face challenges which requires the development of appropriate tools over time.

A further aspect that emerges clearly from the survey is that automation will need to play an important role. Whilst third on the list of challenges, it is also directly related to the first two

identified challenges of rising data volumes and an increase in the time taken to perform an investigation. Automation could certainly help to decrease the workload of digital forensic investigators and help them to have more time for tasks requiring human thinking and less time on routine tasks which could be carried out automatically by scientifically designed algorithms. This would also require the involvement of artificial intelligence in this scenario, wherein tasks of increasing complexity could be automated.

The legislative issues need to be handled by trying to establish mutual understanding, cooperation amongst the nations of the world and cutting across physical and political barriers to have a common agenda and a uniform set of regulations and frameworks for attacking cyber criminals. This would certainly expedite the time frames involved in digital forensic trials and prosecutions. The profiling of cybercriminals would also go a long way to ensure that the time spent to track down cybercriminals is reduced. An elaborate database of offenders and suspects could be used to narrow down the search for potential criminals. This is certainly useful in the case of computer crimes, in which the boundaries of physical barriers are literally absent and could otherwise make the tracking down of suspects quite difficult.

5.10 Conclusion

In the early sections of this chapter, the challenges to the digital forensic field were discussed from both the current and future perspectives. It may be seen that numerous technological innovations that ease the information processing, have also led to difficulties in tracing cybercriminals. Various factors, such as online data storage to the rising use of Cloud Computing add another layer of complexity to digital forensic challenges. The pervasive nature of the Internet which is different from the rigidly defined political and national boundaries is also considered as a major challenge in terms of jurisdictional issues. Also, the lack of uniform resources worldwide is seen as a major resource challenge along with the fact that the amount of suitably trained digital forensic experts is far fewer than the amount required to deal with ever rising numbers of computer crime.

In the second part of this chapter, a survey study of investigating practitioners' and researchers' opinions regarding the current and future challenges of digital forensics was presented. The overall picture that emerges from this analysis is that forensics of all types are

bound to increase in the coming times. Due to the pervasive usage of mobile devices, it is considered that mobile forensics will play a leading and important role. Experts were also concerned about the existing legislation challenges and the need to acquire more cooperative laws amongst the different nations of the World as they expect the current legislations to be inadequate to handle the complexities of modern day cyber related crimes. Also, all individuals in the survey agreed that no available tools for digital forensic investigations actually provide a foolproof solution which can handle all the chores associated with digital forensics.

Most of all, the huge volume of data, the time taken for an investigation and the lack of automation were identified as the top three limitations of existing digital forensic tools. As a result, a tool which can provide automatic analysis will definitely be needed as it can be used to solve those three limitations. Once the analysis phase of the investigation is automated, more data can be processed simultaneously and hence the time taken for an investigation will be reduced accordingly. To this end, the next chapter will focus upon a feasibility study that uses the SOM clustering technique for automatically analysing case artefacts during a digital forensic investigation.

6 Artefact Clustering Using SOM

Based upon the existing research literature and the survey study from the previous chapter both of which focus on the digital forensic tool's capabilities and the associated challenges, it is evident that existing digital forensics tools and techniques still have room for improvement. Also, the survey revealed that there is an acknowledged need to develop an effective approach that can solve these challenges; such a novel approach needs to be a faster and efficient system that features automating the monotonous tasks associated with such investigations.

This chapter aims to research the use of one popular clustering technique - Self Organising Map (SOM) within the digital forensic domain, and to discover the extent to which SOM is capable of providing automatic clustering within the scope of computer forensics. It is anticipated that by the end of the study, a conclusion will be drawn to assess whether SOM can assist investigators with conducting analytical processes more efficiently through searching for patterns within datasets, and producing visual displays of relationships between data.

The chapter also examines and evaluates the clustering process and describes the model itself, as well as discussing prior SOM applications for digital forensic investigations. Moreover, the chapter presents details of a study which investigates the capability of SOM clustering based upon four digital forensic cases.

6.1 Introduction

Cluster analysis or pattern classification involves organising data as a collection of patterns represented as vectors, and is a technique of data mining based on grouping data objects into multiple clusters. The aim is to group objects with higher similarity in one cluster while putting the dissimilar ones in other clusters. The basis of deciding whether objects are similar or dissimilar is the attribute values that are used for describing the objects (Ballabio and Vasighi, 2012).

Clustering analysis is most commonly used as an unsupervised learning technique that has the ability to group data and to find hidden patterns within a given dataset. The clustering algorithms can be utilised to organise and categorise data for various purposes including data compression and model construction. Also, cluster analysis or pattern classifications can be used as both a standalone data mining tool or as a step before applying other data mining algorithms on the data clusters that have been detected (Agarwal et al, 2013). There are several clustering algorithms such as Hierarchical Clustering, K-means Clustering, Hidden Markov Module, and Self-Organizing Map (SOM). Amongst these cluster techniques, it was clearly understood that SOM is not necessarily the most effective approach, but SOM is the most commonly used unsupervised module. Also a ground on using SOM for solving digital forensic related issues has already been established in research. As a result, SOM was chosen as the initial method for exploring the automation on the analysis phase of the digital forensic investigation. Details of SOM and existing studies of SOM within the domain of digital forensics are presented in the following section.

6.2 Self Organising Map (SOM)

SOM was extensively developed by Kohonen during the period from early 70's to late 80's (Bacao and Lobo, 2004) and is recognised as an excellent tool for pattern classification performance. SOM refers to a neural network model that can be used for clustering and visualisation of high dimensional data in a low dimensional space, usually 2D (as illustrated in Figure 6.1) (Kohonen, 2013). It is a technique that is based on unsupervised competitive learning, meaning that the learning process is completely data driven and the neurons or the nodes in the output layer compete amongst themselves. Also, the neurons are connected to the neighbouring ones with a neighbourhood relation. Accuracy and capability of generalisation is dependent on the number of neurons. SOM is extensively used in areas including but not limited to statistical analysis, biomedical analysis, industrial analysis and so forth (Kohonen, 2013).

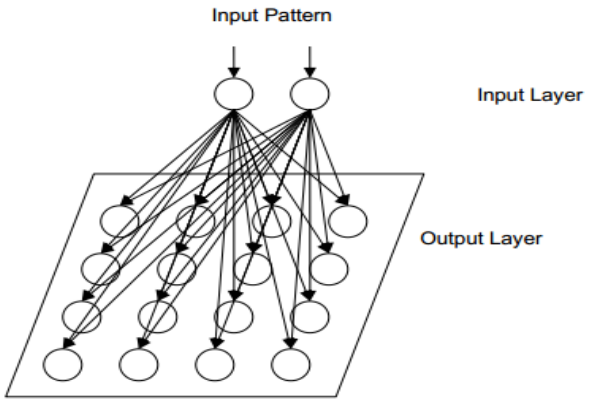

Figure 6.1: The Self-Organising Map (Fei et al, 2006)

6.2.1 Prior Application of SOM to Computer Forensics

Due to its competitive learning nature, SOM can automatically classify the input data without any supervision. Since its invention, SOM has been extensively used in many computer security related fields, including intrusion detection, biometrics and wireless security (Feyereisl and Aickelin, 2009). The use of SOM within the forensic domain can be traced backed to the early 2000s when police started using the technique to link records of serious sexual attacks (Adderley and Musgrove, 2001). Since then, a number of studies have been devised to investigate the suitability of SOM for digital forensic investigations.

Fei et al (2005 and 2006) explored the use of SOM as a support technique to interpret and analyse data generated by computer forensics tools through data visualisation. In their studies, a public dataset containing 2,640 graphical images was utilised; each image contained four features: file name, extension, creation time and creation date. SOM clustered the data after being manually enumerated, producing various two dimensional maps. These visualisations enabled digital evidence examiners to locate information of interest in a more efficient and accurate manner. However, no detailed experimental results were presented to highlight the efficiency and accuracy of their proposed approach.

With the purpose of improving the result of text-based searches, Beebe and Clark (2007) proposed a novel method that utilised SOM to post-retrieval cluster text string search results within an image. In order to test their hypothesis a software tool (named “Grouper”) was developed; Grouper was able to perform a number of activities, including data preparation and SOM clustering. Two datasets were utilised: one was a real-world divorce case and the

other was an artificially created murder case; the image sizes for both cases were 40 and 10 GB respectively. Experimental results demonstrated that the approach can reduce human analytical time by around 80%, despite some additional computer processing time being required (Beebe et al, 2011).

Kayacik and Zincir-Heywood (2006) created a topological model of known attacks for forensic analysis of anomalous network traffic by employing the SOM algorithm. Their model was tested using the KDD 99 intrusion detection datasets. The results of their empirical study show that attacks can be successfully grouped by SOM with a high overall accuracy (i.e. 89.8%). They suggested that the model can be utilised for analysing new attacks or suspicious network behaviour.

Similarly, Palomo et al (2011) focussed upon the analysis and visualisation of network traffic data via the use of SOM, to identify abnormal behaviour or intrusions. For their experiment, a dataset with 150,871 packet samples was created by monitoring a university network via WireShark over a four day period; each sample contained nine features, including IP addresses for source and destination, port numbers, protocol type, time and date stamps, and packet length. The data was clustered using SOM with various network configurations (e.g. 3x3 and 5x5 network sizes). Their experimental results demonstrated that suspicious network traffic was identified by SOM providing vital information for network forensic examiners.

Wang et al (2015) proposed a graphical model to analyse the relationship between criminals through SOM visual analytics. Their model evaluated a dataset with 16,383 features of 16 suspects. Within the model, SOM was used to reduce the features and provide a visual aid to investigators for a better understanding of suspects' activities. According to their experimental results, the proposed model offered assistance and resulted in a more efficient forensics analysis. However, the degree of assistance which the model was able to provide was not clearly reported.

As demonstrated above, SOM has been used in several digital forensic domains, including image analysis, network forensics and text-based searching. The results suggest that SOM can be used successfully to assist forensic examiners in the visualisation of artefacts and the reduction of human analytical time. Nevertheless, the ability to use SOM to analyse an entire

forensic image, whilst being specifically tasked with identifying notables (i.e., those are relevant to the case) or noise (i.e. those are NOT relevant to the case), was never undertaken.

6.3 Experimental Methodology

The purpose of the experiment is to explore the extent to which SOM can be used to cluster artefacts of interest. In the real world, the nature and type of digital forensic cases would vary considerably and it would be expected to face a wide variety of different cases due to the diversity of computer crime instances. The parameters in question include (but are not limited to) image size, number of artefacts, number of relevant files, nature of these files, and even factors such as the qualifications and experience of the digital forensic examiners. Given the practical limitations and feasibility, it was not possible to test the proposed system on a very large number of cases. Yet, representative samples were chosen across a spectrum of possible scenarios. The experiment was configured to repeatedly run each set of the feature input up to five times to achieve results with maximum accuracy. The output from these five repetitions were then summed up and divided by the number of the repetitions (i.e. 5) to calculate the average output; this gave a clear estimated result for each set of experiments. The iteration principle was implemented to investigate the impact of repeating the process to enhance the system efficiency by picking up more notable files related to the case. This reduces the investigation time required to process the case and consequently decide if the subject device contained creditable evidence that could be used in a law suit.

6.3.1 Datasets

Whilst the use of a large number of forensic cases would have provided a stronger evaluation of the approach, access to cases and the time taken to analyse them (to provide the ground truth) were significantly limiting factors. For the purposes of this study, a total of four cases (two public and two private) were observed and analysed. It is important to test any proposed system with a diverse set of inputs in order to judge if it could be implemented in reality. The results are then compared with a similar set of results that have been previously produced during normal forensic analysis (i.e. manually analysed by investigators). It would not be suitable to test the hypothesis using new or unresolved cases since it would be difficult to ascertain whether the results were authentic or in line with the expected findings. Additionally, including different types of cases with different sizes gives an indication of the performance of the proposed approach.

The details of the respective cases are provided in the subsequent paragraphs and highlighted in Table 6.1. As it can be seen the cases vary in image size from Megabytes to hundreds of Gigabytes. Similarly, commonly occurring scenarios in everyday life have been included (i.e. Cyber stalking, blackmailing, ATM frauds and official document fraud). Also, an example of notable/noise files from the Public 1 case is presented in Figure 6.2, where the files marked with 1 are notables and those with 0 are noise.

Case ID	Case Nature	Image Size	Total Artefacts	Total Number of Notables	Notables to the total Artefacts (%)
Public 1	Blackmail/Stalking / Threatening	500 MB	11,638	796	6
Public 2	Hacking	4.5 GB	22,373	11,696	4
Private 1	Official documents fraudulent	16 GB	6,654	30	0.04
Private 2	ATM cards fraud	585 GB	3,456,219	281	0.008

Table 6.1: Sample case details

Creation Time	Modification Time	Path	Type	Notable
14/05/2002 17:31:42	05/06/2002 00:36	102-0283_IMG.J	GIF	1
05/06/2002 01:22:45	05/06/2002 01:22	You Have Been	GIF	0
05/06/2002 01:22:46	05/06/2002 01:22	your chance	GIF	0
03/06/2002 22:12:36	03/06/2002 22:12	Inbox.dbx	GIF	0
03/06/2002 21:15:28	03/06/2002 21:16	Name, Account a	JPEG	1
31/03/2002 17:09:46	14/05/2002 17:57	Dads email	PNG	1

Figure 6.2: An example of notable and noise files from Public Case 1

For the two public cases, the first one was provided by a well reputed digital forensic vendor "Encase" (Guidance Software, 2015) named as "Public 1", with a relatively small image size (i.e. less than half a Gigabyte). The total number of the artefacts was 11,638. The total number of files found relevant to the case was 753, representing an approximate of 6% of the total number of files in the case. The case was a blackmail/stalking incident in which two criminals demand a ransom from a man, otherwise they would harm his daughters. The second public case was supplied by one of the most accredited forensic research and training institutions (NIST, 2014). The case had an image size of 4.5GB and a total number of artefacts of 22,373 files of which 6,179 files were with a creation and access date and timestamp; in addition, 1,001 files of those with a timestamp were found within the case

interest. Notably, the total number of files of interest was 11,669 files which was more than 52% of the total artefacts within the case. The difference between the total artefacts in the case and the detected notables was due to the fact that those artefacts belonged to deleted executables and held no timestamp. It is worth mentioning that these two public cases were both small in size when compared to real and current cases as the estimated hard drive capacity for laptops is one Terabyte (at the time of writing the thesis), with over 25 million artefacts on average.

For the two private real life cases, they were subject to the experiment to prove that the concept works for different types of case and real-life examples. Both of the real cases took place in the Sultanate of Oman and were previously imaged by the legal authorities in Oman CERT. Prior to gaining access to these real cases, a Non Discloser Agreement (NDA) was initiated by the legal authorities (General Prosecution – Oman) and signed by the researcher in order to maintain the confidentiality of the contents, thus preserving the privacy rights of those convicted figures.

The first private case (labelled as Private 1) was an official document fraud case; it was imaged from a flash memory with a size of 16 Gigabytes. The total number of artefacts was 6,654 files, out of which 3,471 files were with a timestamp and 30 files were found to be relevant to the case. The percentage of the files of interest to the total artefacts was 0.04%.

The fourth case was labelled as Private 2, regarding credit card fraud. The image size was 585 Gigabyte and contained nearly 3.5 million artefacts. However, the total number of files with a timestamp was 117,141 and only 281 files were notables. The case was filed upon several complaints received form banking institutions who stated that ATM skimming was occurring in different locations within the same city.

6.3.2 Procedure

All the four cases (both publicly available and privately attained) were processed and analysed by the researcher who is certified as a forensic examiner (AccessData Certified Examiner – ACE) using one of the globally well-known and legally accredited digital forensic tools - "FTK" Forensic Toolkit (AccessData, 2015). The results coming out of this manual examination provided the ground truth to which can be used to evaluate the feasibility of the proposed method which uses SOM to cluster artefacts. In this way, the

accuracy of the SOM analysis process can be calculated. Once the SOM clustered the data, the output in each cluster will be analysed in terms of notables and noise against the ground truth: notables are those are related to the case while the noise are those are not related to the case. The cluster with a high accuracy means as large proportion of notable files are clustered as possible.

During the analysis phase, the case examination was performed as it would be analysed in any forensic laboratory; all the available artefacts in each case were manually checked and accessed to find out what files were relevant to the investigation. Such a process took a long time, especially for the private cases (one contained nearly over 3 million artefacts). The following table presents the estimated analysis time taken for each case;

Case ID	Featured Categories				Total files	Estimated Manual Analysis Time
	File List	Email	Internet	EXIF		
Public 1	✓	✓	✓	✓	11,638	1 day
Public 2	✓	✓	×	×	22,373	1 day
Private 1	✓	×	×	×	6,654	1 day
Private 2	✓	×	×	✓	3,456,219	3 days

Table 6.2: Estimated analysis time for the sample cases

The fundamental process underpinning the use of SOM is capitalizing upon the meta-data associated to the artefacts. Key to this is the file list metadata describing the file, such as its path, creation and accessed timestamps. In addition, it was recognised that metadata existed at the application layer – such as Internet activity, email, EXIF data, Skype, Recycle Bin, are also important. In this research the focus was given to analysing the filelist, email, Internet and EXIF JPEG as they were deemed the most common yet important file categories on a computer system.

Those four categories were exported to a CSV file format and the files of interest in each case were manually marked. The decision on which file category is analysed first is usually made by taking into consideration the nature of the case. For example, if the suspect was accused of hacking, the software and tool that were installed on the seized device would be given a high priority. Also, the investigator would consider an in-depth analysis by delving further in order to not miss any irrefutable or creditable evidence, such as a list of the executable files would

be given high priority within a hacking case. In comparison, if the case was related to child pornography, the multimedia and graphics files would take the highest priority among the selected file categories for analysis. Whilst in a case of monetary fraud, money laundering, bribery or corruption, the investigator would consider financial processing applications and Microsoft Office Excel files or equivalents and so forth.

The analysed cases had at least one of the chosen metadata categories for SOM analysis. The importance of analysing the files under these categories is driven from the fact that rich and important information relating to those metadata categories can be manipulated and gathered for further analysis. For example, whilst a File List SOM might cluster objects based upon a creation time or date stamp, an EXIF SOM would potentially be able to cluster files based upon the GPS location or a camera model. Another important reason to analyse different applications' artefacts is to find and map the associated activities of the suspect, revealing a complete picture and a clear view about what activities have taken in place on the device. It is important to mention that these categories have been selected to test the functionality of the approach, and to determine to what extent the approach would be efficient. Also, more file categories may be selected by the investigator depending on the nature of the case.

The selected features from the chosen categories (the File List and the three applications) were enumerated so they can be processed by SOM. It is worth mentioning that only selected features of those categories were enumerated. For instance, features such as the file extension, the file path, whether the file has been carved, deleted, encrypted or duplicated were selected for enumeration and SOM analysis from the File List. Details of the features for each of the metadata categories are presented in Table 6.3, Table 6.4, Table 6.5, and Table 6.6 representing the File list, email, Internet and EXIF respectively.

Selected Features	Description
Creation date and time	Holds the universal date and time of the document when it was first created
Access date and time	Displays the date and time when the document was last viewed
Modification date/time	Presents the date and time of the last saved changes made to the document
File path	The location of the stored file
File extension	The file name's suffix which specifies the file type
Carved	Those files that have been deleted but extracted from unallocated clusters
Deleted	Files that are temporarily or permanently deleted
Encrypted	Files protected with an encryption tool
Duplicated	Files which have more than one copy

Table 6.3: Selected features for the exported File List category

Selected Features	Description
Subject	Refers to the subject matter of discussion
File Name	Refers to the file name
To	Holds the email address of the recipient
From	Holds the email address of the sender
CC	Carbon Copy of the email addressed to another recipient
BCC	Blind carbon copy addressed to another recipient but hidden from the other recipients
Submit date and time	The date and time that the email was submitted
Delivery date and time	The date and time the email was delivered
Unread	Emails that were received but not yet read by the recipient
Unsent	Drafted emails that have not been sent to the recipient
Has attachment	Emails that contained attachments
Physical -Size	Refers to the physical size of the file
Logical - Size	Refers to the logical size of the file

Table 6.4: Selected features for the exported Email category

Selected Features	Description
Access date and time	Holds the values of the date and time that the website was visited
File Name	Refers to the file name
URL	Presents the website address
Number of hits	Holds values of the visiting statistics for a certain website

Table 6.5: Selected features for the exported Internet category

Selected Features	Description
Last write date	Refers to date the file was last written
Last access date	Refers to the date the file was last accessed
Date taken	Refers to the date the photo was taken
Camera make	Refers to the camera make

Table 6.6: Selected features for the exported JPEG EXIF category

The File List consists of different file categories that are found within the imaged suspect's drive and contains the general information about the different artefacts such as the timestamp, the file path, the file extension and other related information that is required when performing digital forensic analysis. Most of the files that have been generated by the suspect and retrieved by the forensic tool would be present in this list. The following table shows the notable files from all categories subject to the analysis;

Case	Case nature	File List	Email	Internet	JPEG EXIF	Total Notables
Public 1	Blackmail	456	3	65	229	753
Public 2	Hacking	871	29	101	-	1,001
Private 1	Official doc. fraud	30	-	-	-	30
Private 2	ATM fraud	261	-	-	20	281

Table 6.7: Notables files distribution across the various Input Features

As can be seen from the above table, in the case Public 1 the number of notable files presented in the File List was 456 files and the total notable files after analysis of the Email, Internet and JPEG EXIF categories increased to 753 files. Therefore, even those files that are not included in the File List due to the timestamp consideration are still subject to analysis via the metadata that is attainable during the analysis of the other categories. As a result, the number of notable artefacts increases and the case would have the benefit of a fair trial. The same situation was observed for the case Private 2, where the number of notable artefacts also increased after analysing other metadata categories. However, this was not true for case Private 1 where the File List was the only category available for analysis. In comparison, not all features that are found within the selected categories were appropriate; only certain ones were included as they were related to the digital forensic process. For instance, selecting those features relating to the email list would give the investigator a clear picture about the suspect's communication and whether some other individuals are involved in the crime that is subject to investigation. Exporting the EXIF JPEG metadata would provide important information about those images found in the suspect's device such as the timestamp, the camera make and last access date. Furthermore, the Internet File List would provide details about the browsing history of the suspect and would highlight the websites most frequently visited by the suspect. From a research quality aspect, it was imperative to examine the proposed system capabilities with different SOM network sizes against all those cases provided for the experiment, to document the trend and the impact of the network size upon the examined data size.

6.4 Clustering Analysis

All four cases presented in the previous section were subjected to the clustering process using SOM networks with different sizes. The four cases have very different characteristics. The purpose of this analysis was mainly to assess whether SOM is able to perform clustering for digital forensic purposes, and if so, ascertain to what degree the clustering works. In the subsequent sections, the influence of SOM network size and the different input categories are presented and analysed in detail.

6.4.1 SOM Network Size Impact

All the cases were tested with a variety of network sizes, 3x3, 5x5, 7x7, 9x9 and10x10 as it was not obvious which network size would be the ideal for solving the given problem (as is

common in these types of classification problem [Ballabio and Vasighi, 2012]). Using a varied range of network sizes, the influence of the network size on the quality of the clustering can be determined. As a result, the network size was the only design parameter evaluated in this phase of the clustering analysis. The data was analysed by using the density of the notable files versus the noise (i.e. files that do not have any forensic value in the case) within a single cluster. Furthermore, if a majority of the notables are grouped within several clusters, it is very important that the density of notables in these clusters is also as large as possible. In principle, a large density of notable files in the cluster is the main contribution of clustering in improving the productivity of data analysis. Ideally, the density of notable files should be as high as possible within a given cluster. According to the network size setup, a set of experiments was conducted and the analysis of these experiments is presented in following subsections.

6.4.1.1 Clustering Performance for Network Size 3x3

As previously mentioned and highlighted in Table 6.7, not all categories existed in all four cases. Therefore, only the available metadata categories are analysed for a given case. To clearly present the picture, the output data of the SOM network size 3x3 and is consolidated in a unified matrix as illustrated in Table 6.8; also for illustration purposes, only the top 5 clusters according to the density of notables are presented and the remaining notables are grouped in the "Remaining" category

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Public 1	1	405	88.8	186	15.3	3	100	52	100	23	35.4	284	13.5	113	49.3	-	-
	2	39	8.6	187	15.4					22	33.8	207	9.8	95	41.5	-	-
	3	12	2.6	122	10.0					13	20.0	83	3.9	6	2.6	-	-
	4	-	-	-	-					7	10.8	239	11.4	5	2.2	-	-
	5	-	-	-	-									5	2.2	-	-
	Remaining	0	0	720	59.3	0	0	0	0	0	0	1,292	61.4	5	2.2	-	-
Public 2	1	337	38.7	844	18.9	29	100	44	100	42	41.6	39	5.9				
	2	238	27.3	920	20.6					26	25.7	50	7.5				
	3	129	14.8	130	2.9					16	15.8	73	11.0				
	4	64	7.3	363	8.1					15	14.9	75	11.3				
	5	47	5.4	136	3.0					1	1.0	92	13.8				
	Remaining	56	6.5	2,076	46.5	0	0	0	0	1	1	336	50.5				

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Private 1	1	17	56.7	301	8.7												
	2	6	20.0	22	0.6												
	3	3	10.0	228	6.6												
	4	2	6.7	474	13.8												
	5	2	6.7	479	13.9												
	Remaining	0	0	1,937	56.4												
Private 2	1	261	100	19,031	16.3									10	50.0	17	7.5
	2													8	40.0	8	3.5
	3													1	5.0	27	11.9
	4													1	5.0	27	11.9
	5																
	Remaining	0	0	97,849	83.7									0	0	147	65.2

Table 6.8: Clustering output for network size 3x3 (*: the actual number, **:the percentage)

The category File List was presented in all cases, and the quality of the SOM clustering was considered good for the following observations. For Public 1 case, the clustering performance was excellent with 88.8% of the notables found in a single cluster. For case Private 1, the clustering was also quite efficient considering very few notables were contained in this category. Indeed, the majority of the notables (i.e. 86.7%) were found in three out of the nine clusters. For case Private 2, the clustering performance was faultless with all 261 files of notables (100%) concentrated in a single cluster.

The most interesting result that deserves special discussion was obtained from Public 2 case. Looking in more detail at the data for the nine clusters, it could be seen that about 89% of the notables were found in three of the 9 clusters (similar to the performance of Private 1). The cluster with the largest number of notables in this category (337 notables) has a notable density of about 28.5%. The other two clusters with high numbers of notables have densities of 20.5% and 50%. Nonetheless, this still indicates a positive clustering performance in this case as a very large proportion of the notable files in the case was clustered within 3 clusters: a fact that strongly supports the clustering ability and efficiency in this regard.

The email category was present in only two cases, Public 1 and Public 2. For both cases, the clustering was considerably efficient as all notable files are classified in a single cluster; this could be caused by the low number of notable files and few files in total in this category; also the high degree of similarity of files within this category. Therefore, the notable density was not changed with respect to the general population of objects. For instance, there were only 3 and 29 notable files for cases Public 1 and 2 respectively within the email category

The Internet category was also only presented in cases Public 1 and Public 2. Case Public 1 presented fewer notables in this category and they were grouped in four out of the nine clusters (clearly demonstrated by Table 6.8). While, the density of notables in these four clusters is around 10% or less, it is still much better than in the total object population. Similarly for case Public 2, the majority of notables (98%) were distributed within four out of the nine clusters. Also, the density of notables in these four clusters (i.e. 16.7% - 51.9%) is higher than the one in case Public 1.

The EXIF category was also presented in cases Public 1 and Private 2, with case Public 1 having considerably more notables than Private 2 does. Nevertheless, the outputs of the SOM

clustering were very similar for both cases; more than 90% of the notables were grouped within two of the nine clusters. The density rate of case public 1 was 100% for all clusters as all the files were notables; nonetheless, they were all processed as in reality the nature of the file can only be identified after it is examined. In comparison, the density rates for the top two clusters for case Private 2 were 37% and 50%; these figures are good in terms of the notables and noise ratio within this category.

6.4.1.2 Clustering Performance for Network Size 5x5

As illustrated in Table 6.9, the clustering results for the File List category were very good for cases Public 1, Private 2, Private 1 under the network configuration of 5x5; indeed, for these three cases, almost 100% of the notables were grouped with the top five clusters. Also for case Private 2, all the notables were grouped in a single cluster. In terms of the density rate of the notables, they are improving in comparison with the ones in the 3x3 network size. For instance, in Private 2 case, the density rate increased from 1.4% to 5.6%. Also, clusters with almost 100% density rate start to appear, such as two of the top five clusters for case Public 1.

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Public 1	1	175	38.4	66	5.4	3	100	52	100	22	33.8	63	3.0	36	15.7	-	-
	2	106	23.2	2	0.2					15	23.1	148	7.0	20	8.7	-	-
	3	83	18.2	19	1.6					15	23.1	104	4.9	18	7.9	-	-
	4	76	16.7	0	0					13	20.0	83	3.9	17	7.4	-	-
	5	12	2.6	113	9.3									17	7.4	-	-
	Remaining	4	0.9	1,015	83.5	0	0	0	0	0	0	1,707	81.2	121	52.9	-	-
Public 2	1	139	16.0	235	5.3	29	100	44	100	19	18.8	7	1.1				
	2	124	14.2	670	15.0					18	17.8	1	0.2				
	3	108	12.4	84	1.9					16	15.8	48	7.2				
	4	86	9.9	42	0.9					15	14.9	11	1.7				
	5	69	7.9	149	3.3					10	9.9	2	0.3				
	Remaining	345	39.6	3,289	73.6	0	0	0	0	23	22.8	596	89.5				

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Private 1	1	17	56.7	302	8.8												
	2	6	20.0	11	0.3												
	3	3	10.0	227	6.6												
	4	2	6.7	151	4.4												
	5	2	6.7	137	4.0												
	Remaining	0	0	2,613	75.9												
Private 2	1	261	100	4394	3.8									10	50.0	5	2.2
	2													8	40.0	3	1.3
	3													1	5.0	12	5.3
	4													1	5.0	14	6.2
	5																
	Remaining	0	0	112,486	96.2									0	0	192	85

Table 6.9: Clustering output for network size 5x5

For the Email category, all the artefacts including notables and noise were grouped in a single cluster for both Public 1 and 2 cases. This is similar to what was obtained from the network configuration size of 3x3.

Regarding the Internet category, the results were quite different between cases Public 1 and 2. As demonstrated in Table 6.9, all of the notable files were grouped within the top four of 25 clusters for case Public 1. In contrast, only 77.2% of the notables were clustered within the top five clusters for case Public 2. Nevertheless, the density rates for notables of case Public 2 in the top five clusters (two clusters with at least 83.3%) are much higher than its counterpart case Public 1 (25.9%, the best of the top five clusters).

Regarding the EXIF category, it is important to mention that case Public 1 had considerably more notables in this category than case Private 2. In case Private 2, 100% of the notables were grouped within the four clusters (out of the total of 25 clusters): the same as achieved using the network size of 3x3; nevertheless, the performance achieved by network size 5x5 is slightly better than the 3x3 network because the former has larger density rates in the top four clusters than the latter. In comparison, the files were scattered around these 25 clusters for case Public 1. This would be caused by the nature of these files: firstly they were all notables and also they have a very high level of similarity between them.

6.4.1.3 Clustering Performance for Network Size 7x7

As illustrated in Table 6.10, the best performance for the File list was obtained on Private case 2 as all the notables were concentrated in a single cluster; also the noise level is reduced in comparison with the results from network sizes 3x3 and 5x5. The performance for cases Public 1 and Private 1 are also good as the notables present within the top five clusters covered more than 64% of the notables within the cases. In comparison, the results from case Public 2 only achieved a grouping of 46.7% of the total notables within the top five clusters (despite this the result itself is reasonable). In terms of density rate, case Public 1 obtained the best result as three of the top five clusters contained 100% of notables; while the remaining two clusters (out of the top five) had at least 81% of the notables density rate, showing the SOM clustering works well.

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Public 1	1	100	21.9	19	1.6	3	100	52	100	22	33.8	52	2.5	23	10.0	-	-
	2	58	12.7	0	0.0					14	21.5	81	3.8	20	8.7	-	-
	3	51	11.2	12	1.0					13	20.0	83	3.9	17	7.4	-	-
	4	44	9.6	0	0.0					8	12.3	71	3.4	13	5.7	-	-
	5	43	9.4	0	0.0					7	10.8	81	3.8	13	5.7	-	-
	Remaining	160	35.2	1,184	97.4	0	0	0	0	1	1.6	1,737	82.6	143	62.5	-	-
Public 2	1	114	13.1	93	2.1	29	100	44	100	16	15.8	19	2.9				
	2	86	9.9	42	0.9					15	14.9	1	0.2				
	3	74	8.5	488	10.9					14	13.9	5	0.8				
	4	69	7.9	122	2.7					13	12.9	4	0.6				
	5	64	7.3	146	3.3					10	9.9	2	0.3				
	Remaining	464	53.3	3,578	80.1	0	0	0	0	33	32.6	634	95.2				

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Private 1	1	13	43.3	113	3.3												
	2	6	20.0	11	0.3												
	3	3	10.0	131	3.8												
	4	2	6.7	126	3.7												
	5	2	6.7	159	4.6												
	Remaining	4	13.3	2,901	84.3												
Private 2	1	261	100	2631	2.3									8	40.0	3	1.3
	2													7	35.0	0	0.0
	3													3	15.0	5	2.2
	4													1	5.0	1	0.4
	5													1	5.0	7	3.1
	Remaining	0	0	114,249	97.7									0	0	210	93

Table 6.10: Clustering output for network size 7x7

As previously mentioned, the email category was present in only two cases; Public 1 and Public 2. The same observation was obtained from a 7x7 network when compared with network sizes 3x3 and 5x5. All the notables and noise files were placed in a single cluster.

The results for the Internet category were quite different between the two relevant cases, Public 1 and Public 2. Based upon the information presented in Table 6.10, for Public 1 case, it is obvious that almost 100% (i.e. 98.4%) of the notables of this category were contained within the top five clusters of the total 49 clusters. In comparison, for case Public 2 67.4% of the notables were found within the top five clusters: almost 30% less than those present for case Public 1. Nevertheless, regarding the density of the notables within the top five clusters, case Public 2 outperformed case Public 1 with at least 73.7% of the notable density rates being observed within the top proportion of clusters, while the notable density rates presented in case Public 1's top five clusters are much lower.

Considering the EXIF category under the network configuration size of 7x7, for case Public 2 all the notables were grouped within the top five clusters as shown by Table 6.10 Also, the density rates are getting better in comparison with the results obtained from the 5x5 configuration. Within the top five clusters, one cluster with 100% and two clusters with 50% and above density rate for the notables. In comparison, the performance gets worse for case Public 1 as the total notables are scattered within more clusters.

6.4.1.4 Clustering Performance for Network Size 9x9

Results similar to those observed from the network size 7x7 configuration were obtained from the network size 9x9 setup, for the cases Private 1 and Private 2's file list category. In both cases, the performance was very good as 93.3 % and 100% of the notables were grouped within the top five clusters for cases Private 1 and 2 respectively (as demonstrated in Table 6.11). In addition, the noise level reduces by 6.7% from the network size of 7x7 for case Private 1, consequently increasing the notables' density rate. Regarding the performance for cases Public 1 and 2, a reasonable level of performance was obtained as over 37% of the notables were grouped within the top five clusters (out of the total 81 clusters; far better than random guessing). In addition, the notable density rates are pretty good for these two cases. Especially for the case Public 1 where at least 82% of the notable density rate was achieved by the top five clusters.

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Public 1	1	47	10.3	0	0.0	3	100	40	76.9	15	23.1	23	1.1	17	7.4	-	-
	2	41	9.0	9	0.7					15	23.1	58	2.8	10	4.4	-	-
	3	39	8.6	4	0.3					9	13.8	39	1.9	10	4.4	-	-
	4	36	7.9	0	0.0					7	10.8	40	1.9	10	4.4	-	-
	5	32	7.0	4	0.3					6	9.2	18	0.9	8	3.5	-	-
	Remaining	261	57.2	1,198	98.7	0	0	12	23.1	13	20	1,927	91.4	174	75.9	-	-
Public 2	1	86	9.9	42	0.9	29	100	44	100	16	15.8	19	2.9				
	2	73	8.4	28	0.6					11	10.9	1	0.2				
	3	61	7.0	6	0.1					10	9.9	1	0.2				
	4	58	6.7	214	4.8					10	9.9	0	0.0				
	5	47	5.4	136	3.0					10	9.9	2	0.3				
	Remaining	546	62.6	4,043	90.6	0	0	0	0	44	43.6	642	96.4				

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Private 1	1	13	43.3	112	3.3												
	2	6	20.0	6	0.2												
	3	4	13.3	75	2.2												
	4	3	10.0	59	1.7												
	5	2	6.7	55	1.6												
	Remaining	2	6.7	3,134	91												
Private 2	1	261	100	2,171	1.9									8	40.0	3	1.3
	2													7	35.0	0	0.0
	3													3	15.0	4	1.8
	4													1	5.0	1	0.4
	5													1	5.0	5	2.2
	Remaining	0	0	114,709	98.1									0	0	213	94.3

Table 6.11: Clustering output for network size 9x9

Regarding the email metadata category, the same observation for case Public 2 was obtained as all files (both noise and notables) were classified in a single cluster. However, for case Public 1, as seen in Table 6.11, all the notables were grouped in a single cluster apart from the noise files. As a result, this improved the density of notables and thus the clustering performance in comparison with the previous SOM network setups (i.e. 3x3, 5x5 and 7x7).

For the Internet category, both cases (Public 1 and 2) achieved good performance. In terms of the proportion of notables that were grouped within the top five clusters, Pubic 1 case achieved better results than case Public 2 did: 80% of the notables file were grouped within the top five clusters for case Public 1 while only 56.4% for case Public 2. Regarding the density of notable files, case Public 2 got a better result: at least 83.3% of the notable density rate was obtained in four of the top five clusters, demonstrating the effectiveness of the SOM clustering technique.

As previously stated, the EXIF category was only available in the cases Public 1 and Private 2, with case Public 1 having considerably more notables in this category than the Private 2 case did. Similarly to the other network setups, the files from case Public 1 were processed despite all of them being notables. Regarding case Private 2, the performance was good as 100% of the notables were grouped within the top five clusters (out of the total of 81 clusters) with only 5.7% of the noise grouped within the same clusters. Also, high notable density rates were observed from these results: over 37.5% of the density rates are achieved in four of the five top clusters.

6.4.1.5 Clustering Performance for Network Size 10x10

The approach was further examined with the maximum size of network for a more complete investigation. As demonstrated in Table 6.12, for the File List category of case Private 2, all notables were grouped within a single cluster, similar to what was obtained using other SOM network size setups (i.e. 3x3, 5x5, 7x7 and 9x9); in addition, the purity of the cluster improved as less noise files were grouped within the selected cluster. For the other three cases, despite the proportion of overall notable files within the top five clusters being lower than those obtained from previous SOM network configurations, the results are still very promising e.g. 38.6% of the notables were grouped within the top five clusters for case Public 1. Also, the density of notables within the top five clusters of all the three remaining cases is larger in comparison with previous SOM network setups.

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Public 1	1	50	11.0	12	1.0	3	100	38	73.1	15	23.1	21	1.0	21	9.2	-	-
	2	37	8.1	0	0.0					15	23.1	30	1.4	9	3.9	-	-
	3	32	7.0	1	0.1					8	12.3	21	1.0	8	3.5	-	-
	4	31	6.8	0	0.0					7	10.8	42	2.0	7	3.1	-	-
	5	26	5.7	2	0.2					6	9.2	6	0.3	7	3.1	-	-
	Remaining	280	61.4	1,200	98.7	0	0	14	26.9	14	21.5	1,985	94.3	177	77.2	-	-
Public 2	1	86	9.9	42	0.9	29	100	44	100	16	15.8	0	0.0				
	2	52	6.0	203	4.5					12	11.9	0	0.0				
	3	50	5.7	3	0.1					10	9.9	2	0.3				
	4	47	5.4	136	3.0					9	8.9	0	0.0				
	5	47	5.4	78	1.7					9	8.9	0	0.0				
	Remaining	589	67.6	4,007	89.8	0	0	0	0	45	44.6	663	99.7				

Category		File List				Email				Internet				EXIF			
		Notable		Noise		Notable		Noise		Notable		Noise		Notable		Noise	
Case ID	Top Cluster ID	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%	A*	%**	A	%
Private 1	1	7	23.3	33	1.0												
	2	6	20.0	79	2.3												
	3	6	20.0	11	0.3												
	4	2	6.7	20	0.6												
	5	2	6.7	33	1.0												
	Remaining	7	23.3	3,265	94.8												
Private 2	1	261	100	1,864	1.6									8	40.0	0	0.0
	2													6	30.0	3	1.3
	3													2	10.0	0	0.0
	4													2	10.0	5	2.2
	5													1	5.0	1	0.4
	Remaining	0	0	115,016	98.4									1	5	217	96.1

Table 6.12: Clustering output for network size 10x10

Regarding the email metadata category, the results for both cases (Public 1 and 2) are similar to what those obtained before: i.e. all the notables were grouped within a single cluster for case Public 1 and all the files (both notables and noise) were grouped in one cluster for case Public 2. For case Public 1, the proportion of noise files within the selected cluster is reduced to 73.1% of the total noise files.

For the Internet metadata category, it is very clear that more than 45% of the notable files were clustered within the top five clusters (out of the total 100 clusters) for both the Public 1 and 2 cases. Also the amount of noise files within these clusters is smaller than those from previous network setups; as a result, the density of the notable files within these clusters also increased; for example, four out of the five clusters were with a 100% density rate for notables from case Public 2.

For the EXIF metadata category, the performance for case Public 1 gets worse as the notables were scattered around more clusters due to more clusters being offered. In comparison, the SOM clustering performance for case Private 2 is very good: 95% of the total notables were grouped within the top five clusters and the density rates of these notables within these clusters are also high: two clusters with 100%, one with 66.7% and one with 50%. These demonstrate that the SOM clustering works well on the given problem.

6.4.2 SOM Input Categories Impact

The previous analysis focused exclusively on the influence of network size on the SOM clustering performance. It was demonstrated that the performance differs for the different cases even for the same network size. Thus, in this section, the input categories are analysed in terms of the total number of artefacts with timestamps, total number of notables and percentage of notables. Also, these parameters are different for each of the analysed cases in terms of numbers and percentages. The analysis was conducted separately for each input feature, as not each category was present in all cases.

The File List category was present in all of the four analysed cases as it contains all files within a forensic image. However, there were large differences in the values of the parameters that were considered for the analysis from all four cases. Table 6.13 represents the additional parameters that are taken into the account for the analysis, in order to address the impact of file list metadata as an input feature.

Case ID	Total analysed files	Tot. Notables w/timestamp	Notables(%)
Public 1	1,671	456	27.2
Public 2	5,340	871	16
Private 1	3,471	30	0.08
Private 2	117,141	261	0.002

Table 6.13: Notable files from File List in all cases

As can be seen, the table shows a significant difference between the first three cases with an approximate range between 1,000 and 6,000 artefacts of the file list, and the last case with more than 100,000 artefacts with timestamps. Also the number of notables is widely different from as few as 30 in case Private 1 to 871 in case Public 2. Finally, the percentage of notables in the total number of artefacts of this category varies from 0.002% for case Private 2 to 27.2% for case Public 1. The clearest distinction between the cases can be made with respect to the percentage of notables, with Public 1 and Public 2 both having a relatively high percentage of notables in comparison with cases Private 1 and Private 2 which have very low percentages of notables of under 0.1%.

As illustrated in Table 6.14, in three of the four cases, clustering based upon the file list alone proved very successful. For cases Public 1, Private 1 and 2, all notable files were obtained within five clusters with at least 56.4% irrelevant files clustered in the remaining clusters when using the 3x3 SOM configuration. Indeed, case Public 1 was able to identify 100% of notables (within three clusters) with 59.3% of the irrelevant files being grouped in the other six clusters. While Private 2, identified 100% of the notables in a single cluster and only introduced 1.6% of the noise. In comparison, only case Public 2 did not identify all notable files – it was able to cluster 93.5% of notables at a cost of including 53.5% of the irrelevant files – still resulting in a huge reduction in the number of files an investigator would need to analyse if these five clusters were successfully identified during an investigation. The worst result was given by the 10x10 network in case Public 2: only 32.4% of the notables were collected within the chosen five clusters, with the remaining 67.6% scattered around the other 95 clusters.

Network Size		3x3		5x5		7x7		9x9		10x10	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
Public 1	1	88.8	15.3	38.4	5.4	21.9	1.6	10.3	0.0	11.0	1.0
	2	8.6	15.4	23.2	0.2	12.7	0.0	9.0	0.7	8.1	0.0
	3	2.6	10.0	18.2	1.6	11.2	1.0	8.6	0.3	7.0	0.1
	4	-	-	16.7	0	9.6	0.0	7.9	0.0	6.8	0.0
	5	-	-	2.6	9.3	9.4	0.0	7.0	0.3	5.7	0.2
	*	0	59.3	0.9	83.5	35.2	97.4	57.2	98.7	61.4	98.7
Public 2	1	38.7	18.9	16.0	5.3	13.1	2.1	9.9	0.9	9.9	0.9
	2	27.3	20.6	14.2	15.0	9.9	0.9	8.4	0.6	6.0	4.5
	3	14.8	2.9	12.4	1.9	8.5	10.9	7.0	0.1	5.7	0.1
	4	7.3	8.1	9.9	0.9	7.9	2.7	6.7	4.8	5.4	3.0
	5	5.4	3.0	7.9	3.3	7.3	3.3	5.4	3.0	5.4	1.7
	*	6.5	46.5	39.6	73.6	53.3	80.1	62.6	90.6	67.6	89.8
Private 1	1	56.7	8.7	56.7	8.8	43.3	3.3	43.3	3.3	23.3	1.0
	2	20.0	0.6	20.0	0.3	20.0	0.3	20.0	0.2	20.0	2.3
	3	10.0	6.6	10.0	6.6	10.0	3.8	13.3	2.2	20.0	0.3
	4	6.7	13.8	6.7	4.4	6.7	3.7	10.0	1.7	6.7	0.6
	5	6.7	13.9	6.7	4.0	6.7	4.6	6.7	1.6	6.7	1.0
	*	0	56.4	0	75.9	13.3	84.3	6.7	91	23.3	94.8
Private 2	1	100	16.3	100	3.8	100	2.3	100	1.9	100	1.6
	*	0	83.7	0	96.2	0	97.7	0	98.1	0	98.4

Table 6.14: Experimental results for the File List category of the four cases

The Email category was present in only two of the analysed cases (Public 1 and Public 2) and the number of notables was relatively low in both cases. Table 6.15 represents the additional parameters that are taken into account for the analysis, in order to address the impact of the email list as an input feature:

Case ID	Total emails	Notable emails	Notables (%)
Public 1	55	3	5.45
Public 2	73	29	39.72

Table 6.15: Notable % for email files from cases Public 1 & Public 2

As the table shows, there is a large difference in the percentage of notables with more than 34% for the 2 Public cases in regards to this category, indicating that the performance was better. However, the total number of artefacts as well as the number of notables in this category is relatively low compared to the File List category. As illustrated by Table 6.16, all the artefacts (both notables and noise) were grouped in one cluster for both cases; although around one quarter of the benign artefacts were separated from the notables for the network sizes 81 and 100 for case Public 1. Reasons for this phenomenon could be due to the small amount of total email artefacts (53 and 73 for cases Public 1 and 2 respectively), or a high

level of similarities that were presented within them.

Network Size		3x3		5x5		7x7		9x9		10x10	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
Public 1	1	100	100	100	100	100	100	100	76.9	100	73.1
	*	0	0	0	0	0	0	0	23.1	0	26.9
Public 2	1	100	100	100	100	100	100	100	100	100	100
	*	0	0	0	0	0	0	0	0	0	0

Table 6.16: Email clustering for cases Public 1 and 2

The Internet category was present in the cases Public 1 and Public 2, with the case Public 1 presenting fewer artefacts in the category than case Public 2, and the effect of the network size is different between the two cases. The following table represents the total number of artefacts versus the notable ones and the contributing percentage for each.

Case ID	Total Internet files	Notable files	Notables (%)
Public 1	2,170	65	2.99
Public 2	766	101	13.18

Table 6.17: Percentage of Internet Notables for the cases Public 1 & Public 2

The results for cases Public 1 and 2 are presented inTable 6.18. More than 75% and 50% of the notables were grouped within the chosen five clusters for cases Public 1 and 2 respectively. The best performance (in terms of the proportion of notables) was obtained by using the network size 5x5 for case Public 1: 100% of the notables were distributed in four clusters with only 18.8% of the total noise being clustered within the same clusters. For case Public 2, the worst performance was achieved under the configuration of the 10x10 network size: only 55.4% of the notables were successfully clustered by the SOM within the chosen five clusters; however, due to the high density of notables within each cluster (four with 100% of notables and one with 83.3% of notables), merely two noise artefacts (i.e**.** 0.3% of total noise in the case) were classified within those five clusters.

Network Size		3x3		5x5		7x7		9x9		10x10	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
Public 1	1	35.4	13.5	33.8	3.0	33.8	2.5	23.1	1.1	23.1	1.0
	2	33.8	9.8	23.1	7.0	21.5	3.8	23.1	2.8	23.1	1.4
	3	20.0	3.9	23.1	4.9	20.0	3.9	13.8	1.9	12.3	1.0
	4	10.8	11.4	20.0	3.9	12.3	3.4	10.8	1.9	10.8	2.0
	5	-	-	-	-	10.8	3.8	9.2	0.9	9.2	0.3
	*	0	61.4	0	81.2	1.6	82.6	20	91.4	21.5	94.3
Public 2	1	41.6	5.9	18.8	1.1	15.8	2.9	15.8	2.9	15.8	0.0
	2	25.7	7.5	17.8	0.2	14.9	0.2	10.9	0.2	11.9	0.0
	3	15.8	11.0	15.8	7.2	13.9	0.8	9.9	0.2	9.9	0.3
	4	14.9	11.3	14.9	1.7	12.9	0.6	9.9	0.0	8.9	0.0
	5	1.0	13.8	9.9	0.3	9.9	0.3	9.9	0.3	8.9	0.0
	*	1	50.5	22.8	89.5	32.6	95.2	43.6	96.4	44.6	99.7

Table 6.18 :Experimental results for the Internet category of cases Public 1 and 2

The EXIF category was found in two cases: Public 1 and Private 2. However, in case Public 1 all the EXIF files with timestamps in the suspect's data were notables, thus they were all identified and flagged during the clustering process across all network size values. The total EXIF files, the notable files, and the percentage of those notable files are the key parameters considered for analysis purposes, and are listed as follows:

Case ID	Total EXIF files	Notable files	Notables (%)
Public 1	229	229	100
Private 2	246	20	8.13

Table 6.19: Percentage of EXIF Notables for the cases Public 1 and Private 2

As illustrated by Table 6.19, all the EXIF files within Public 1 were notables; the results also highlight that the SOM is capable of sorting data according to their similarities. In contrast with the results presented by Private 2, a better set of outcomes are observed as 90.8% of the notable files were grouped within two clusters when using the network size 3x3 configuration. Regarding Private 2, more than 95% of the notables can be found within the chosen five clusters for all network configurations; also the proportion of noise within the these five clusters reduces significantly as the network size increases: 34.8% of noise for network 3x3 in comparison with only 3.9% for network 10x10. Moreover, 100% of the notable density rate can be observed under the network sizes 7x7, 9x9 and 10x10, reinforcing that SOM can be used for clustering information with very high performance.

Network Size		3x3		5x5		7x7		9x9		10x10	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
Public 1	1	49.3	-	15.7	-	10.0	-	7.4	-	9.2	-
	2	41.5	-	8.7	-	8.7	-	4.4	-	3.9	-
	3	2.6	-	7.9	-	7.4	-	4.4	-	3.5	-
	4	2.2	-	7.4	-	5.7	-	4.4	-	3.1	-
	5	2.2	-	7.4	-	5.7	-	3.5	-	3.1	-
	*	2.2	-	52.9	-	62.5	-	75.9	-	77.2	-
Private 2	1	50	7.5	50	2.2	40	1.3	40	1.3	40	0
	2	40	3.5	40	1.3	35	0	35	0	30.0	1.3
	3	5	11.9	5	5.3	15	2.2	15	1.8	10	0
	4	5	11.9	5	6.2	5	0.4	5	0.4	10	2.2
	5	-	-	-	-	5	3.1	5.0	2.2	5	0.4
	*	0	65.2	0	85	0	93	0	94.3	5	96.1

Table 6.20: Experimental results for the EXIF category of Cases Public 1 and Private 2

6.5 Discussion

Based upon the results presented in tables in the last two subsections, the application of SOM upon artefact identification appears to work very well. Indeed, for three (i.e. File List, Email and EXIF) of the four chosen categories through all four cases, more than 93.5% of notables were grouped within the top five clusters at least under one SOM network configuration, with at least half of the irrelevant files not being included. The best performance of all results (in terms of grouping most notables within the top five clusters and also minimising the proportion of noise) was achieved by using the 10x10 network for the File List of case Private 2: all the notables were clustered in a single cell with only 1.6% of the total noise being present. If this were utilised by digital evidence examiners, the amount of their workload could be dramatically reduced; providing an opportunity for more cases to be processed within a given time.

In comparison, under each category and each network configuration, the worst performance was obtained within case Public 2. Further examination of the case reveals a large proportion of the artefacts were carved executables – as such they did not appear in the File List or application-level metadata. It is clear, for such an approach to work successfully the artefacts (or a significant proportion of them) need to appear within the metadata outputs. In most cases this is possible, for example these carved executable files actually might have metadata located in the Recycle Bin (INFO2 record); however, more application-level metadata needs to be processed.

In general, a larger proportion of notables can be obtained within the top five clusters by using smaller SOM network sizes (e.g. 3x3 or 5x5); however, a considerable amount of

noisy/irrelevant files were also observed. A reduction in the noise can be achieved by choosing larger SOM network sizes (e.g. 9x9 or 10x10) with a compromise of the number of notables (as demonstrated in Figure 6.3, Figure 6.4, Figure 6.5 and Figure 6.6). Also, when the SOM network size increases, more clusters with a higher density rate of notables start to appear. Therefore, the granularity of the results is proportional to the sizes of the SOM network: smaller network sizes provide for coarser grained results while larger network sizes for finer grained outputs. This in part is obvious due to the larger number of clusters available. The need to be able to determine which network size to utilise is likely to be driven by the aim of the investigator (e.g. to obtain all notables at the expense of picking up more noise or obtain some notables to confirm the drive has relevant content at the expense of very little noise).

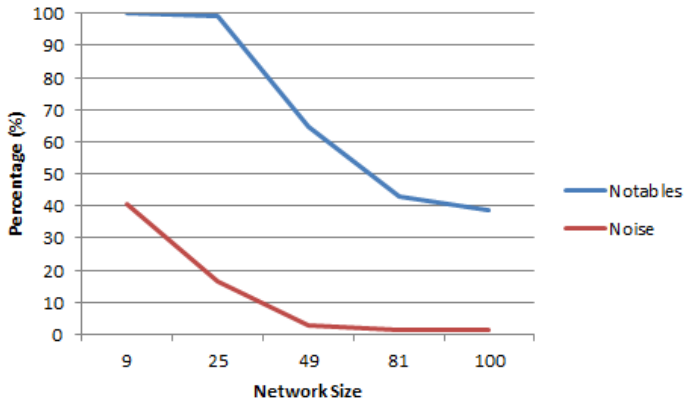

Figure 6.3: Comparison on Notables and Noise in top 5 clusters of the Case Public 1 File List on various network configurations

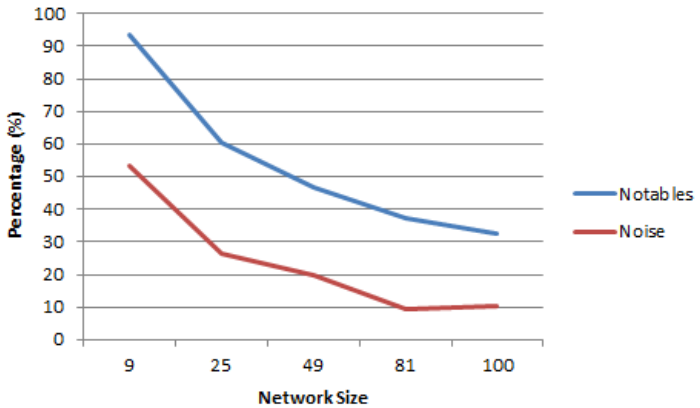

Figure 6.4: Comparison on Notables and Noise in top 5 clusters of the Case Public 2 File List on various network configurations

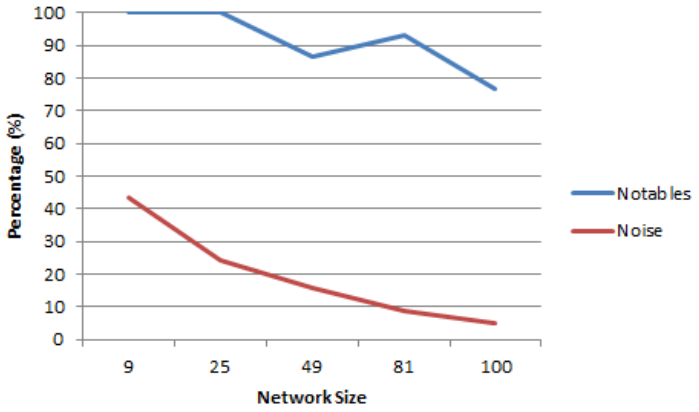

Figure 6.5: Comparison on Notables and Noise in top 5 clusters of the Case Private 1 File List on various network configurations

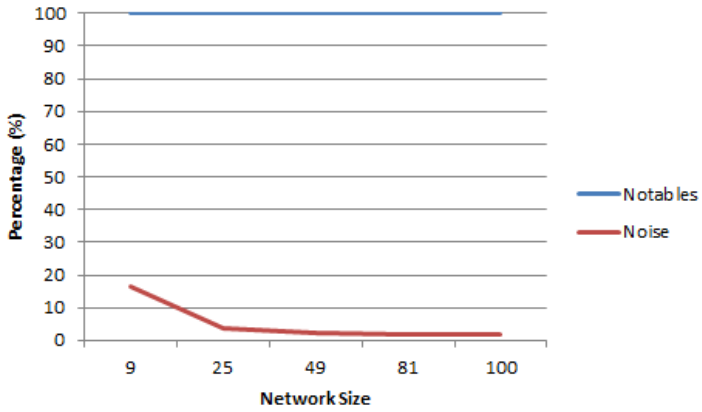

Figure 6.6: Comparison on Notables and Noise in top 5 clusters of the Case Private 2 File List on various network configurations

6.6 Conclusion

A detailed analysis on the SOM clustering performance for various network sizes and different input features was presented. The analysis proceeded first by considering solely the influence of the network size. Then, the analysis considered explicitly the influence of other parameters such as the total number of artefacts, total number of notables and percentage of notables. For each case, the analysis was performed separately for each category of artefact. The analysis of the result shows that clustering works in most analysed cases in the following sense: notables are grouped in a few clusters; and, the density of notables in these clusters is higher than in the general object population. Indeed, in a number of occasions, clusters consisted exclusively of notables, which is the ideal situation. In many cases, the concentration of notables in the clusters was very good, higher than 50%; and in most cases, the concentration was 10 times larger than in the general object population. Only in a few

isolated cases, the density was not improved. Most notably for the email category, in which most results show no improvement of the density after clustering, and this could be one of the consequences when a relatively small number of files are subjected to clustering.

It is fair to say that in most cases the clustering performance was relatively good and also the increase of the SOM network size had mixed effects on the clustering performance. Nevertheless, increasing the network size improved in a few cases as the density of notables in the clusters improved; while in other cases there was no noticeable improvement. As the network size increases (hence larger number of clusters), there is a natural tendency that more clusters would contain notables and this happened in some of the cases. However, the density of the notables in these clusters remained relatively high which indicates a high clustering performance.

Concerning the result's analysis in terms of the given features impact, based on the available data, it can be concluded that a low percentage of notables lead to better clustering performance, although the clustering performance itself depended on the type of artefact considered. However, for the main artefact category, i.e. the File List, that was present in all the considered cases, this conclusion is quite well supported by the available data.

As demonstrated above, SOM can definitely be utilised for the analysis of digital forensic images because of its capability of clustering notables in an accurate and timely fashion. In order to take the full advantage of SOM, a scientific approach for identifying the first and subsequent clusters that contain high proportions of notable files is required; and such an approach is fully discussed in the next chapter.

7 Automated Evidence Profiler (AEP)

According to the positive result from the SOM clustering experiment, it is imperative to design an algorithm that could capitalise upon the output of the SOM clustering technique. For the purpose, a timeline analysis is utilised along the clustering result. Such a novel approach aims to bring both clustered SOMs and timeline analysis together to serve the digital forensic field in a positive manner.

7.1 Introduction

There are a number of reasons as to why an algorithm needs to be developed for the Automated Evidence Profiler (AEP). The main three are as follows. Firstly, the first cluster of notables must be successfully and accurately identified. Secondly, a basis for identifying subsequent clusters across all SOMs (file list and application levels) needs to be present. Finally, the algorithm elegantly provides a refinement process to identify all other relevant artefacts with a reasonable success rate.

In order to determine how to identify the first cluster, an intelligence led approach is utlised. The criteria is fed into the system by the digital investigator to direct the system capabilities towards clustering, locating and grouping those file of interest in related clusters. Given the fact that certain files categories are associated to certain crimes, this would allow the approach to concentrate and focus on certain file categories. For instance, if the investigated case was related to hacking or web defacement, it is highly expected to find hacking related tools that were installed and used in the suspect's device. Similarly, in a case of child pornography, images, videos, or chat logs are the most common files and it is common once someone views or downloads an image, it is most likely to view other related images, thus giving the system the ability to pick up a large proportion of files of interest.

7.2 The Automated Evidence Profiler

The basic fundamental concepts of cyber profiling are based on the premise that common factors exist within computer crimes and cyber criminals. For example, 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 modeling 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 computer crime 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 computer crimes and cyber criminals (Horsman et al, 2011). One of the earliest attempts to build frameworks to tackle computer crimes and bring cybercriminals to justice based on the identification of common factors was the "The Computer crime Execution Stack". This framework was intended to enable the convention of conservative model between the legal formalities and the technical factors (Hunton, 2009).

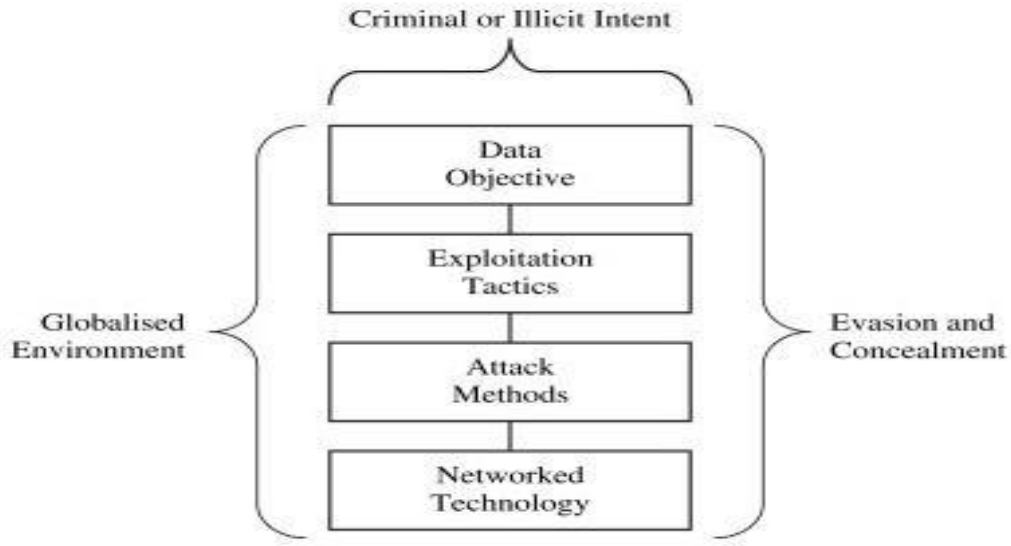

Figure 7.1: The Computer crime Execution Stack (Hunton, 2009)

However, this model was not intended to provide the investigator with technical details for the possible involved individuals or groups, but to formalise the investigation process when dealing with computer crime. Although criminal profiling being implemented in the area of criminal psychology, it is not yet used within the domain of digital forensics. Indeed, little research has been conducted on investigating the links between high-level criminal features and low-level computing-based objects. In addition, the need for computer crime profiling has become an essential due to the continuing rise of computer crimes around the globe.

Nevertheless, some computer crimes such as those against the childhood need to be highly considered and a mechanism to detect those involved in cyber paedophilia activities is imperative to develop. Another example that assures the need for cybercriminal profiling is the abduction cases where more than individual or groups could be involved and the time is critical to save a victim when his life is threatened and could be lost if the demands were not answered.

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 experiments 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.

Figure 7.2: Automated Profiling through Evidence Trails

As illustrated in Figure 7.2, 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 notable artefacts. In this manner, it will be possible to build up an understanding of actions a user undertakes.

The entire concept of AEP would help to reconstruct the crime scene and the events that

occurred before and after. The first component namely the "crime to artefact mapping" is used to map specific types of crime to certain types of files. Artefact mapping is performed through an automated process which checks the nature of artefacts base on previously fed information in the knowledge base (Intelligence led), check outs if there is a high probability of the artefact being related to the case. The US department of Justice categorised the digital crimes and provided an extensive guide for the digital investigators on what to find and were according to the crime category (DoJ, 2001). The Table 7.1 illustrates the mostly likely file types found with the association of specific nature of computer crimes;

Files / Crimes	Browsing History	Emails	Executables	Excel	Graphics	Multimedia	PDF/ Word
Bribery / Money Laundry	-	✓	-	✓	-	-	-
Child Pornography	✓	✓	✓	-	✓	✓	-
Fraud	-	✓	-	-	✓	✓	✓
Hacking	-	✓	✓	-	-	-	-
Illegal Downloading	✓	-	✓	-	-	✓	✓
Illegal Drug Trade	✓	✓	-	✓	-	-	-
Intellectual Property Theft	✓	✓	-	-	-	-	✓
Ransom	-	✓	-	-	-	-	✓
Stalking	✓	✓	-	-	✓	✓	-

Table 7.1: Crimes nature common file types

Considerably, mapping crime to artefacts is the initial automating process in which the final system is built on. The purpose of such a process is to automate the process of mapping the relevant artefacts within the forensic image subject to the analysis. This process, as the other components within the AEP do, aims to highlight the area of interest with limited human intervention, thus saving investigator's time and effort. With the presence of such feature, the AEP is able to deal with different types and nature of computer crimes involving thousands or possibly millions of artefacts that would require a long time to process and analyse.

Evidence trail generation is performed next by using the abstracted data that is clustered into different group. As a result, the relevant activities would be mapped, correlated using the timestamp and this forms the base for the timeline analysis.

Once the Evidence Trail task is executed, the profiling process takes place by Profiling, Filtering and Refinement entity. This profiling process within the system involves studying the user behaviour in order to create a profile template base on the routine activities chronicled with the support of timeline analysis. The filtering process then comes to the line where the collected data is sorted in order to segregate the relevant data from irrelevant, thus identifying the notable artefacts. It is worth mentioning that the filtering process implements a smart approach in which it is able to differentiate between the analysed files, even though they were of same file types through examine several parameters that are pre-set by the investigator such as the timeframe and the file types. For instance, in a case where images and videos are the main concerned file types and millions of images were found in analaysed device, the system would exam the found images and videos against pre given criteria. To be more precise, the Metadata related to the analysed files plays a major role in identifying the relevant files as the last-write, download timestamp, and camera make and owner would differ in those files, thus alerting the system to trigger towards the right target. If a single cycle of the above processes was not sufficient enough to have the desired amount of data to build up the profile, the process would be repeated until it is in a satisfactory state.

The final stage within the AEP lifecycle is the Refining in which the processed artefacts are purified of the noise, prioritized and visualized to the investigator. Correspondingly, the data base and the other interacted components within the AEP are updated to serve the investigator with the latest data in future cases, which in turn would reduce the time taken to analyse a case and the human effort at once.

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.

As illustrated in Figure 7.3, once initial artefacts have been identified through the crime-mapping to artefacts, the AEP automatically creates a series of chronology trails of the artefact – each chronology based upon a context within which it was used (e.g. within the file system, 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 within 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 trails and the purpose of the AEP 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.

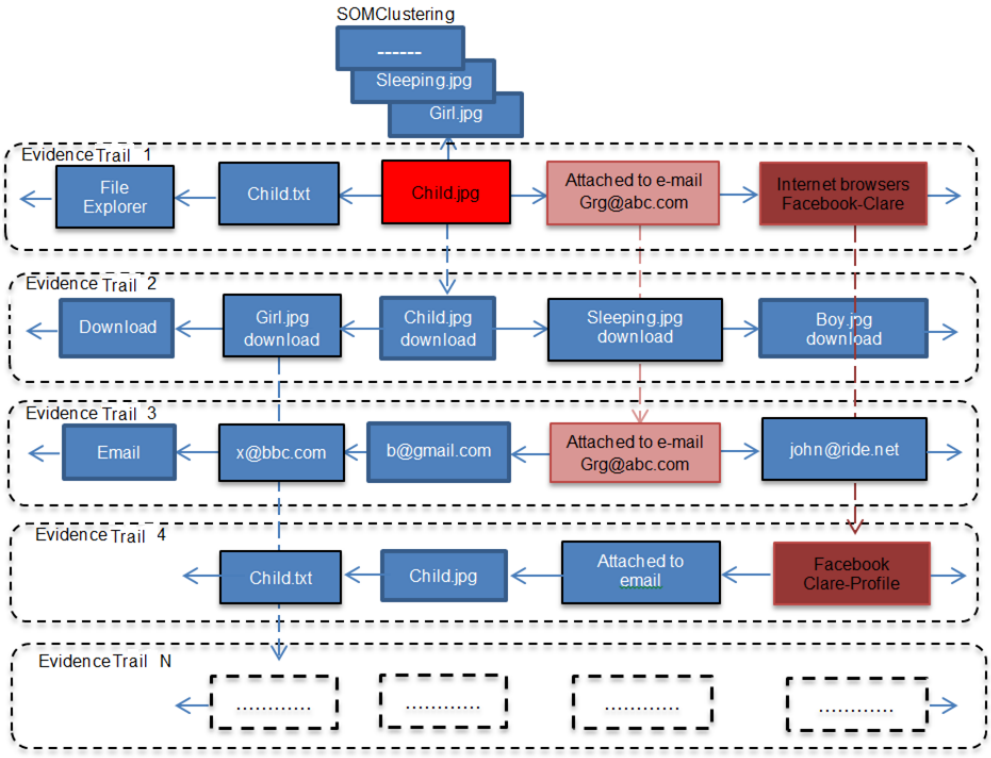

Figure 7.3: Example of Evidence Trails

Once the first cluster is identified and analysed a further analysis step will be performed in order to identify the potential files of interest that have been grouped in other clusters. This step referred to as "Timeline analysis" and aims to link the objects that are subject to the analysis with another objects that have relevant artefacts i.e. window size or timeframe, in this way, this key process develops the concept of Evidence Trails. The TimeLine analysis approach has been implemented as an effective technique to provide the digital investigator with a better understanding about the case that is subject to the investigation. However, it was used as an independent method, not incorporated with an automated digital forensic tool (i.e. a case management tool).

Figure 7.4 illustrates selected file categories that are subject to the analysis within a given timeframe range Three times are given where the 21.00 is the event time in which an

identified file of interest has been created or accessed; while the other two times (20.45 & 21.15) represent the length of the timeframe that has been set by the investigator in which the timeline analysis process will operate.

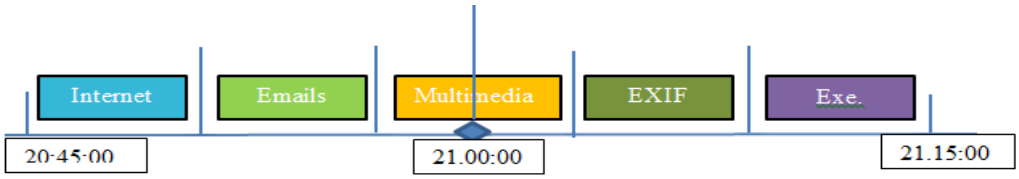

Figure 7.4: Timeframe process

Timeline Analysis is a useful technique that helps creating a chronological record of the events that preceded the digital crime. This helps the digital forensic examiners as well as the legal authorities to have an overview of the series of activities that led to the crime and hence drawing a full picture of the event subject to the investigation. Moreover, this mechanism would be able to highlight the parties or the individuals involved in the crime if they were more than one. Importantly, having the right cluster been identified by the crime profiling process, the key role for the timeline analysis algorithm is to identify the second cluster to analyse. Thus, this process will ensure that only clusters within file and applications level containing notable artefacts are included and incorporated. The fundamental assumption of the timestamp is that when human error occurs, it is likely to be one among many. Therefore, this idea is being conceptualised to timeline analysis as a forensic method of identification.

This may seem to be a simplistic approach; but it is a novel idea since none of the current forensic tools have addressed such an important feature. Although timeline analysis is not relatively new technique, it had been neglected as an effective approach and recently it attacks more attentions.

There are different resources from where the data for traditional timeline can be collected and aggregated from, such as the date and time stamps of the creation, access, and modification upon a file. Different schools of thoughts argued about the usability of traditional timestamps in the digital forensics field. It has been argued that the timestamps can be tampered with; hence deliberately trying to mislead the digital forensic investigators. In comparison, considering the fact that some cases may contain a large proportion of artefacts without timestamp, locating those notable files with timestamp would be easier and quicker using this mechanism.

Also, the examination outs both public and real cases presented in the last chapter, less instances where such an act was performed by the cyber criminals in the real world situations. One of the reasons could be that timestamps exist in various sources such as files system, browsing history, events log file, application used etc. and it is challenging to modify all of them. Another reason could be that large number of files are stored on a computing device ranging from hundreds of gigabytes to terabytes and it may not be practical to tamper them.

Irrespective of the reason for the above, the fact remains that in most of the e-crimes the timestamp is not tampered and therefore the timeline created from these timestamps could provide a clear picture about what had the suspect been doing, and if more than an individual have been involved. The information available from this resource is also extensive since these timestamps are generated automatically and are always present unless they had been changed deliberately. Given these limitations, it can be stated that the timestamp information is useful source to rely on as a timeline analysis approach, especially in the absence of other sources.

7.3 Experimental Methodology

Whilst the approach initiates with the timestamps, the idea of including other SOMs is to build upon that. Thus, the analysis by no means fully depends on the timestamp approach to deal with the artefacts. Furthermore, rather than using the ordinary timestamp, the approach uses additional metadata of the other SOMs. Therefore, the first identified cluster that has a large portion of certain file type related to the crime nature is subject to further investigation. This process is known as "Intelligence led" and it takes place after the expert feeds the system with an input criteria prior to analysis according to the case nature. Figure 7.5 illustrates the process flow starting from a selected commercial forensic tool's (FTK) output:

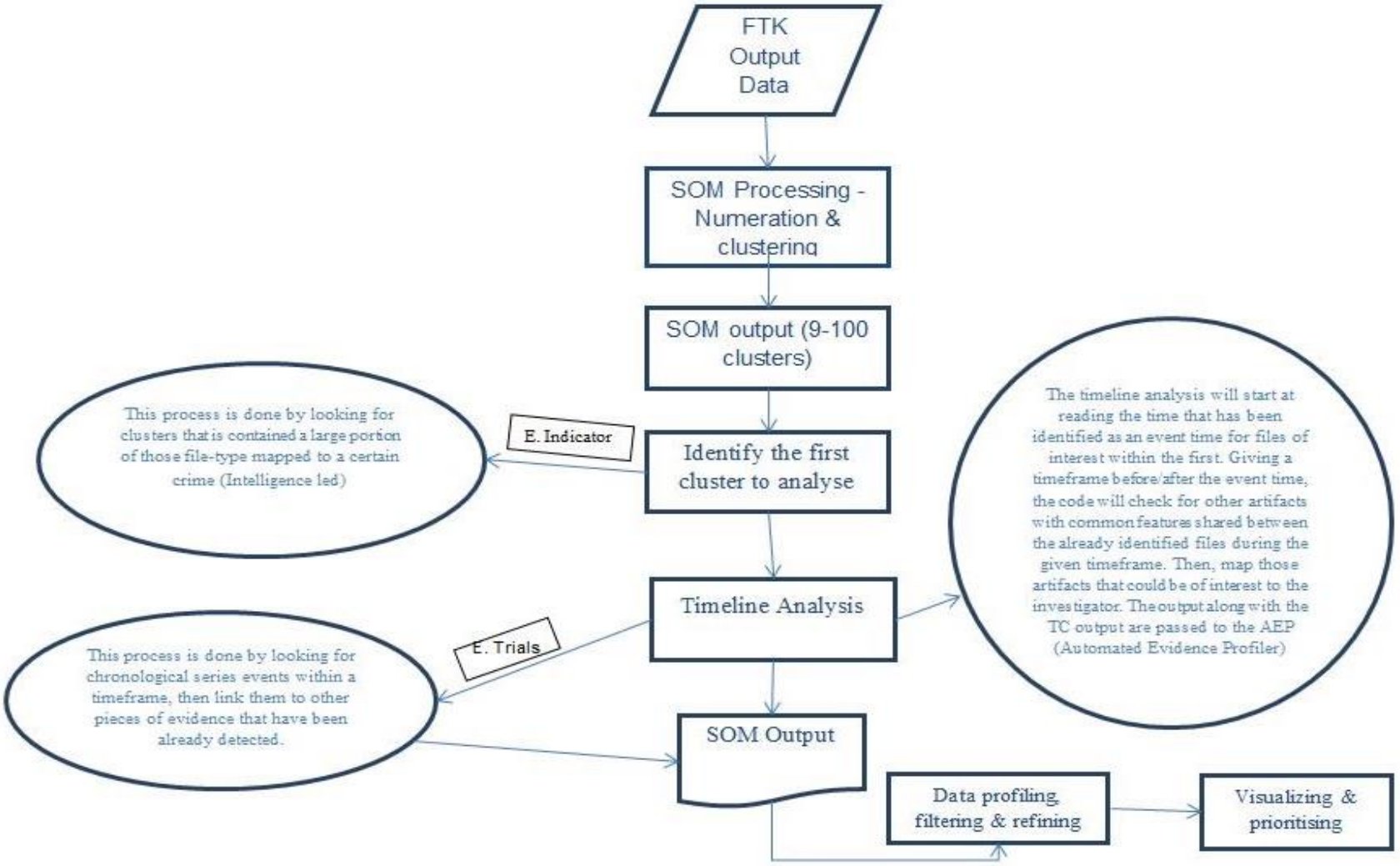

Figure 7.5: FTK to Matlab Process flow

In order to investigate the FTK's output, several variables were implemented to perform various tasks, such as identifying the first cluster, performing timeline analysis, identifying the next cluster to analyse, and so forth. The key variables are listed in Table 7.2.

Feature/Value/ Term	Details
Exp ID	Refers to the case ID
Net Size	Refers to the SOM network size use to examine the data
Min	Refers to the timeframe used to examine the data with
Repeat 1,2,3,4,5	Refers to the repeating time of examine the data using the same features (network size, timeframe, iteration number)
Average	Calculate the average percentage of the result of each iteration
Notables	Represents the percentage of notables which are found via timeline analysis divided by the total notable files in the case
Noise	Represents the percentage of irrelevant files which are found via timeline analysis divided by the total noisy files in the case
All result new (all files)	Represents the percentage of files which are found via timeline analysis (including both Notables and Noise) divided by the total number of files in the case
First cluster	Used to identify the first cluster to analyse based upon the expert advice. It outputs total notable files in the identified cluster, total files (including noise), percentage of the notable files, , percentage of files from expert advice and the ID of the identified cluster

Table 7.2: Key variables lists

To prove the concept, four outputs (Filelist, Email, Internet and EXIF) were produced. The process starts with understanding the crime profile that provides an indication on the type of artefacts that should be focused. The output of this process is initially used to find the first cluster within the Filelist SOM output. Once the first cluster is identified, a timeline analysis with be performed upon every file presented within the cluster to create the evidence trails of files. Then the output will contain a new set of files with associated SOM IDs; these files were accessed during that timeframe on which the analysis is carried out. They will be further analysed and the cluster with more average file occurrence is then prioritised for further analysis.

The average file occurrence is calculated by using the sum of file occurrence divided by the total by number of files (as demonstrated in Figure 7.6). The output of this step that is the average value is used to pick the next SOM cluster which needs to be analaysed, and the cluster with the highest average is chosen irrespective of the total number of files. For instance (as demonstrated in the Figure), the average values of the two clusters are calculated as 3.6 and 16.5 for clusters SOM1 and SOM2 respectively and hence SOM2 is chosen for

further analysis. Also, in case the chosen cluster has already been analaysed, the cluster with the second highest average file occurrence will be selected. In the above diagram,

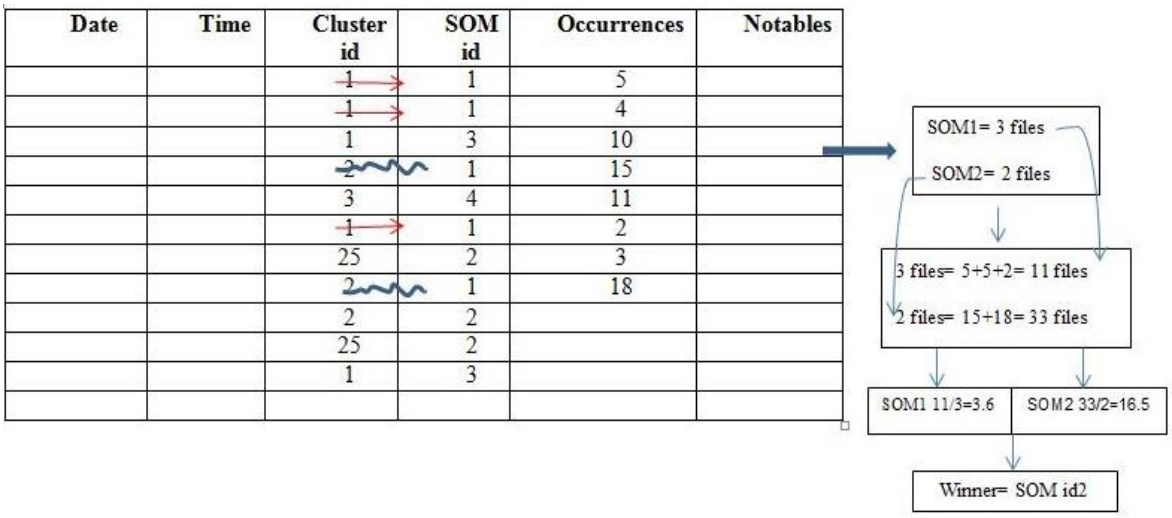

Figure 7.6: Process of timeline analysis

A detailed of how the process of timeline analysis works is given below. As illustrated in Figure 7.6, various inputs, including Timestamp, cluster ID, SOM ID, Occurrences and Notables are presented within the given table. The red coloured horizontal lines indicates those cluster ID 1 that have artefacts from SOM ID 1 while the curly navy lines indicate the next cluster that also contained artefacts belong to SOM ID 1 which in the above scenario was the Cluster ID 2. The next step involves the finding the number of occurrences in each of these SOM clusters and the number of these occurrences have been recorded in the fifth column of the table with the heading "occurrences".

It is also clear from the text boxes along with the table that the 3 files identified in cluster ID 1 essentially are formed with a combination of 11 (i.e. 5 + 4 +2) files while those in cluster ID 2 are formed from a combination of 15 + 18 = 33 files.

The next step involves calculating the mathematical average of these two by dividing the total number of occurrences in a specific SOM ID with the number of files in that. In this case the average values come out to be 3.6 and 16.5 for SOM 1 cluster 1 and SOM 1 cluster 2 respectively.

As the rule states, the winning SOM is picked based on the cluster that contained a higher value of the average and not the actual number of files or occurrences. Hence, the decision on the next cluster to analyse is based on the cluster that contained the higher average of artefacts which in the given scenario are belong to SOM1 cluster ID2. Importantly, as the number of files to be analysed increases, SOM needs more power in order to carry out its increasing task load.

The above steps are constantly repeated until satisfactory results are obtained. It can be clearly seen that the proposed dynamic and iterative approach gives a respectively accurate results which can be relied upon by the digital forensic investigators. With each step of the analysis, the database is refined by AEP. The relevant files of interest are sorted in order of priority and are visually presented to the digital forensic investigators along with the Technical Competency measure output. Thus, the cyclic process of selection based on predetermined set of parameters helps to filter out the relevant files and narrow down the selection in a continuous loop fashion; therefore, a refined list of relevant artefacts with a smaller size in comparison with the original dataset will be generated

Conclusively, the final list generated from the series of iterations would increase those artefacts were created from different categories by the suspect's activity. The result beside other information collected from further analysis, helps the examiners to build up a complete picture of the suspect's activities and to find out whether or not more individuals are involved in the case intentions. As a result, the case can be prepared in a creditable manner from both technical and legal perspectives. Furthermore, such an approach would not require a certified forensic examiner to process e-crime cases which in turn would result in bridging the gap in the global scarcity of digital forensic examiners to process the ever increasing e-crime cases.

7.4 Experimental Results

Upon the implementation of the AEP algorithm, it was imperative to investigate and asses the functionality and the efficiency of the approach with an appropriate input data. The following sections detail the key parameters and the settings of the experiment and provide rich analysis base on certain aspects as detailed in section 7.4.1.

7.4.1 General setting of the experiments

As the main purpose of these experiments was to assess the influence of the three main parameters of the proposed algorithm: network size of the SOM, the timeframe used in the timeline analysis and the number of iterations required to identify the notable files. The network size was varied over a set of five values: 9, 25, 49, 81 and 100. For each of these values of the network size, the time frame length was varied over the following eight values: 0.5, 1, 2, 5, 10, 20, 30, 45 minutes. The varying range of the selected network size values and the timeframe were given as it was not obvious to the researcher what network size and timeframe would be the ideal for a given a case. Hence, for each case, a number of 40 (8x5) experiments were conducted corresponding to all the possible combinations of network sizes and time frame length. This choice allows obtaining a good estimate of the influence that these two important parameters have on the performance of the proposed algorithm, both independently and in conjunction with each other. Also there is a maximum of 4 iterations on finding additional clusters for each experiment. Moreover, each experiment was repeated four times for each given timeframe and network size for all four experimented cases with the average given for those four repeats to judge on the accuracy of the result and consisted in four iterations.

The analysis in this chapter will follow the following aspects:

- The effect of the iterations: how the results vary across the iterations, and how many iterations of the performed iterations were effective.
- The effect of the timeframe: how the results vary across the different sets of timeframes given to measure the impact
- The influence of network size and time frame length on the performance of the algorithm, both separately and jointly;

The analysis is carried out over the four cases (Public 1, Public 2, Private 1 and Private 2) with a special emphasis on the common trends that can be identified from the results obtained for all these cases.

The average value for the different features outputs was considered in three cases, namely Public 1, Public 2, and the case Private 1. However, due to the time limitation, for case Private 2, the given input features were examined for one run only, thus there was no average

value presented. Furthermore, the iteration influence was considered, and its impact on the result was respectively considered.

7.4.2 Performance analysis

The performance is characterized primarily by the number of selected notable and noise files and these parameters of the AEP may vary depending upon the chosen network size, time frame length and the number of iteration applied. The performance analysis performed in this section is aimed at characterizing the influence of these parameters on the performance of the AEP approach, expressed in terms of the percentage of notable and noise files. Table 7.3 gives an overview of the experiment result for the four examined cases and highlights both the proportions of notable and noise files across different network size and various timeframes.

Case ID	Notable (%)	Noise (%)	Network	Timeframe (min)
Public 1	69	25.8	25	0.5
	72.7	29.9	25	1
	81.6	39.6	25	2
Public 2	68.8	21	81	0.5
	75.8	39.6	49	1
	76.4	42	49	0.5
Private 1	83	14	49	20
	84	25.6	25	2
	86	29	25	5
Private 2	92.8	51.5	100	0.5
	92.8	51.8	81	0.5
	92.8	73	25	0.5

Table 7.3: Notable vs Noise rates across different network size and timeframes

7.4.3 Iterations

As far as this factor was concern to be a potential influencing aspect, it was important to study and analyse the result in order to determine if such an influence exists; and if so up to what extent the result can be impacted. It was noted that the percentage of identified notable files increases or stays the same if a high percentages were picked up from the first or second iteration. This is true for all cases. Also, Table 7.4 gives an example for the iteration impact on both notables and noise files across all cases.

Case ID	Notable % per iterations				Noise % per iteration				Network Size	Timeframe (min)
	1	2	3	4	1	2	3	4		
Public 1	28.6	34.9	49.9	52.7	11	13.9	18	22.6	49	0.5
	53	56.5	58.8	61	15	17	18	20	81	0.5
	37.8	53	56	69.	15.9	20	23.6	35	25	0.5
Public 2	79.4	79.8	80.2	80.5	39.4	51.9	53.4	54.1	49	2
	70.8	74.7	77.2	78.5	22.3	29	41.3	51.4	81	2
	45	63	63	63	1	25	35.9	38	81	1
Private 1	23	76.7	83	83	0	4	8	14	49	20
	23	76.7	83	83	0	4	5	7	81	10
	23	76.7	83	83	0	5	5	7	81	30
Private 2	92.8	92.8	92.8	92.8	53.8	59	62	68	49	0.5
	96.4	96.4	96.4	96.4	78	82	82	86	9	0.5
	92.8	92.8	92.8	92.8	42.7	43	45.9	51.5	100	0.5

Table 7.4: The impact of the iteration factor across different network size values and timeframes

There are some differences in the performance where the iteration has less influence upon the result. For example, in the case Private 2, as illustrated on Table 7.4, the notable files percentages did not have considerable changes over the iterations under the SOM network sizes 9, 49 and 100; importantly, in comparison with the public cases, all notable files in this case were driven from the File List and EXIF metadata files only, which may justify such an occurrence of the percentage of files in each iteration. Furthermore, the percentages of identified notable files were considerably high from the first iteration; such an output was expected and did not have a negative impact on the performance degree. In contrast, the percentage of noise files tends to increase, but with less growing rate than the notable files do. Thus the output positively indicates the functionality and the usability of the approach.

To conclude the influence of iteration, it can be seen that the performance varied between cases, depending upon the input data and the amount of notable files contained within. The variety of file categories related to those notables was also a determining factor. Practically, an investigator would not have knowledge of such information unless the case is subject to investigation. It is for this reason that the iteration factor is required in each case.

7.4.4 Influence of Timeframe Length

To assess the timeframe impact on the input data, it was imperative to examine if such an impact exists and to determine the degree to which such an influence has on all examined cases. It is logical that the longer the time length, the more files would be selected during the

time line analysis; thus a high percentage rate pertaining both notable and noise files is expected. In general, this trend is respected by the experimental data.

Table 7.5 shows sample of the timeframe impact on the result for all examined cases. For the case Public1, network size 5x5 was taken as example to examine the impact and different timeframes were selected. It is noticeable that the percentage of notables does increase with a larger timeframe; and it is true for most of the cases. It can be seen from the result listed on Table 7.5, the percentage of the notable files at the timeframe 0.5 minute during the fourth iteration of the case Public 1 was 69%; and the performance increased to 81% under the timeframe 2 minutes for the same case under same SOM configuration. Giving the fact that the notable percentage was relatively high, such an increase was reasonable. As for the noise files proportion, the timeframe also had an impact. However, the increase on the noise files percentage by a longer timeframe was not as high the notables do, resulting in a positive performance as the proportion of the notable to the noise files was respectably high with a rate of at least 40%.

Case ID	Notable % per iterations				Noise % per iteration				Network Size	Timeframe (min)
	1	2	3	4	1	2	3	4		
Public 1	66	68	69	69	23	25	25.8	25.8	5x5	0.5
	70	71.9	72	72.7	28	29	29.9	29.9	5x5	1
	68.9	70.5	75	81	29	30	34.6	39.6	5x5	2
Public 2	72	72	72	76	24.7	28	28.5	42	7x7	0.5
	75	75.8	75.8	75.8	29	34	36	39.6	7x7	1
	79.5	79.6	79.6	79.9	39.5	50	50.5	50.9	7x7	2
Private 1	23	77	77	84	0	10	16	25	5x5	2
	23	78	78	84	0	10	17.5	29.6	5x5	20
	23	78	78	85	0	11	18	29	5x5	45
Private 2	92.8	92.8	92.8	92.8	42	43	45	51	10x10	0.5
	92.8	92.8	92.8	92.8	43	43	58	60	10x10	1
	92.8	92.8	92.8	92.8	70	75	75	76	10x10	2

Table 7.5: The impact of the timeframe factor across different network size values across all cases

As demonstrated the timeframe factor had a significant influence on most of the cases; and, a smaller impact was observed if the given timeframe was short. For instance, for the case Private 2, the average percentage for the notable files did not change over the very short timeframe (i.e. 0.5, 1 and 2 minutes). Considering the fact that a high proportion of notable files were identified, such an output was expected. In contrast, the percentage rate of the noise files was considerably high across all network sizes; nevertheless, such an output did

not have a significant negative result as the percentage rate for the notable files was very high with over 90% across all the network size values.

Considering the case Public 2, the timeframe did now show a significant influence upon both notable and noise files. The section 7.4.5 presents the findings in details.

For the case Private 1, the algorithm performed excellently as a very high notable percentage (84%) was collected across various timeframes (e.g. 2, 20 and 45 minutes). Moreover, the percentage of the noise files had a very minor increase with less than 4% between the shortest and the longest timeframe. The same phenomenon applies to the case Private 2 as a very high percentage of notable files was detected; however a considerable big increase in the noise files percentage was presented as well. Considering the very low number of the notable files found in the case (281 files) compared to the analysed files (117,141 files), such an outcome was expected as the deviation between the both types was considerably massive.

In term of the noise files, the influence of timeframe was noticeable in a way that when larger timeframes were applied the increase in the noisy files was less than its counterpart for the notable files.. However, the extent of such influence varied from case to case due to the case's data size, the data source and the amount of notable artefacts within each case. The best performance among all cases was noticed in the case Public 1 when a combination of relatively small network sizes and short timeframes (ranging from 0.5 to 2 minutes): the percentage of noise files was very low in comparison with notable files (i.e. 39% of the noise files against 81 % of the notables for longest timeframe within the same parameters).

To conclude this section, the extent of the timeframe impact on the noise files varied. The longer timeframe is the more notable and noise files detected. Such an impact was true under most of the varied network size across most of the examined cases. Nevertheless, the increase of notable files plays a key factor on deciding which timeframe should be applied to maximum the gap between both notables and noises with the notables' favour.

7.4.5 Exceptions from the General Trend

In the preceding sections, the impact of the both key factors (Timeframe & Network size) upon the results has been discussed. However, the analysis determined that there are some

exceptions from the general trend have taken place. In the subsequent sections, a detailed analysis illustrates those exceptions is provided.

As for the time frame influence, whilst the output results of the experimented cases had similar patterns in general, exceptions from such trend took place in several situations. Such exceptions normally expected when experimenting data differ in size and type with different features vary from one case to another. Additionally, the influencing factors such as the timeframe, have significantly contributed to those exceptions. In fact, those exceptions provided a boarder view of what would be expected when examining data with different input features under large varieties of circumstances. Notable, those expectations were minors and only noticeable for the case Public 2.

Whilst the timeframe had little influence over the notable files in the case Public 2 with the use of small network size, it also had a low impact on the noise files with an average increase of 10% between the lowest and the highest timeframes. The Table 7.6 presents results with the implementation of network size 3x3 for case Public 2. Sample values were selected to present the impact of the timeframe on the case for demonstration purpose.

Notable % per Iteration				Noise % per Iteration				Timeframe (min)
1	2	3	4	1	2	3	4	
94.9	95.5	96.4	98.0	79.0	80.4	81.6	88.3	0.5
98.5	98.5	98.7	98.7	99.4	99.4	99.4	99.4	10
98.7	98.7	98.7	98.7	99.5	99.5	99.5	99.5	45

Table 7.6: Exceptions from the general trend for timeframe influence from case Public 2 under network size 3x3

It can be seen from Table 7.6 that the percentage of the notable files did not change and remained almost constantly from the shortest to the longest timeframes. Giving the fact that this case was made up for training and development purpose, most of the relevant artefacts consisted of hacking and anti-forensics tools that most of them were deleted and were not processed by the proposed method. Such a fact had a negative impact on the functionality of the timeframe mechanism as those deleted artefacts did not exist in the file system nor on the application level; as a result, the associated metadata were not available and it was not possible for the algorithm to process them.

Regarding the influence of the network size, the time frame length is fixed and the percentages of notables and noise files are represented for the different network sizes that

were used in the experiment. The general trend is that percentages of notables and noise files increase with the increase of the network size – which is logical given the increase in the availability of clusters. As an example that illustrates perfectly this assertion, the following figures represent the notable and noise file percentages for time frame length 0.5 min and different values of the network sizes for case Public 2 - the only case with a noticeable exception (as illustrated in Table 7.7). For all four iterations, as the network size increases, the percentages of notables and noise files decrease. However, the decrease on noises is much more than those on the notables, providing a better result in term of notable to noise ratio. Indeed, when using the combination of the larger network size and the short timeframe, a maximum margin between notables and noise was obtained almost 40% as illustrated in Table 7.7.

Notable % per Iteration				**Noise % per Iteration**				**Network size**
1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	
94.9	95.5	96.4	98	79	80.4	81.6	88.3	3x3
72.7	81.1	88.7	90.7	25.5	52.3	70	74.7	5x5
69	69.1	69.5	78.9	23.3	28.3	28.5	49.1	7x7
65.2	67.2	68.4	68.7	15.5	22.6	26.6	28.9	9x9
54.6	62.2	64.9	64.9	11.6	22.5	22.5	27.6	10x10

Table 7.7: Exceptions from the general trend for the influence of network size for case Public 2 and time frame length 0.5 min

7.4.6 Choice of the Best Performance for the Algorithm

The results presented in the previous section suggest that it is possible to choose the time frame length such that the ratio notables/noise is maximized for each given network size. In this section, the analysis is further pursued to determine if it is possible to draw a general conclusion based on the experimental data that is available for the four cases. For this purpose, the following figures are provided to illustrate best performance for the algorithm for the analysed cases;

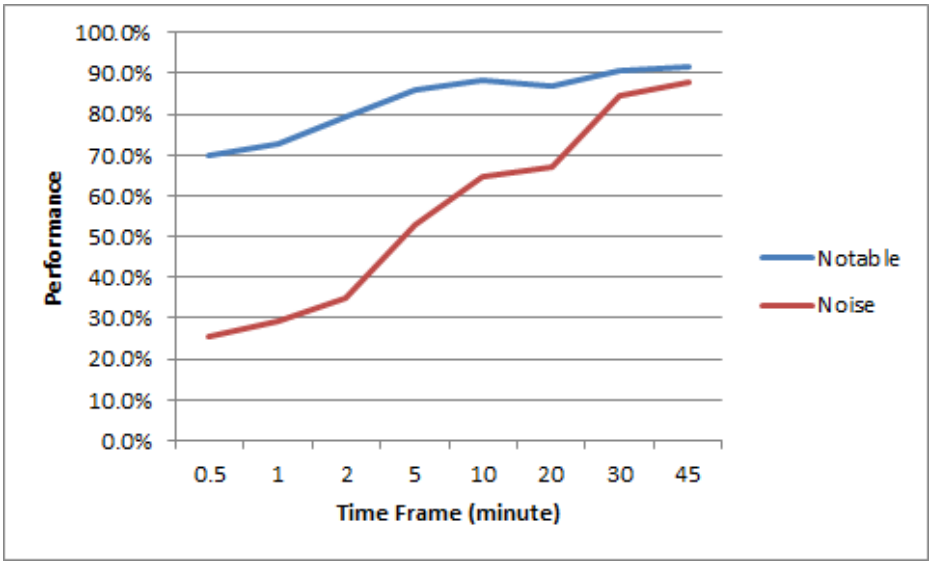

Figure 7.7: Best performance for the Case Public 1- network size 5 x 5

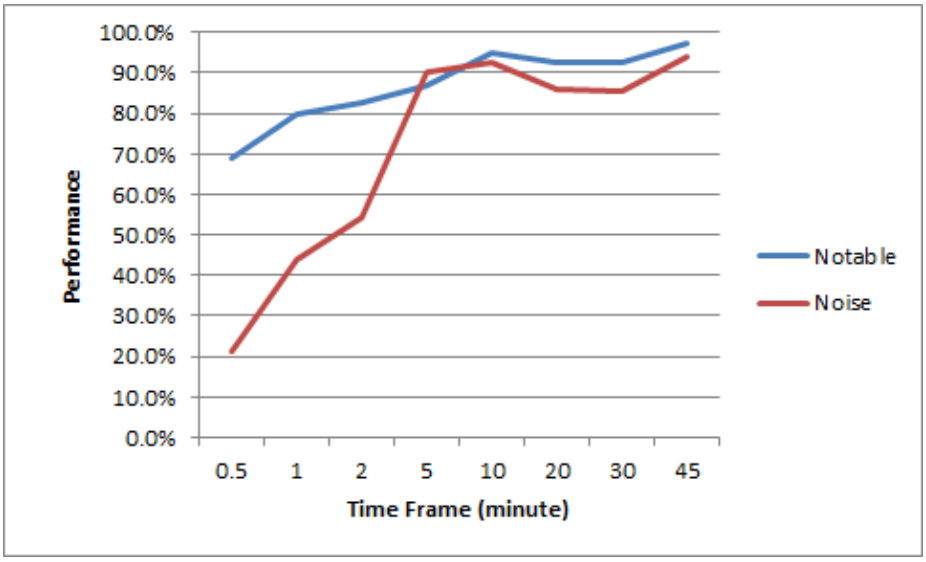

Figure 7.8: Best performance for the Case Public 2 - network size 9x9

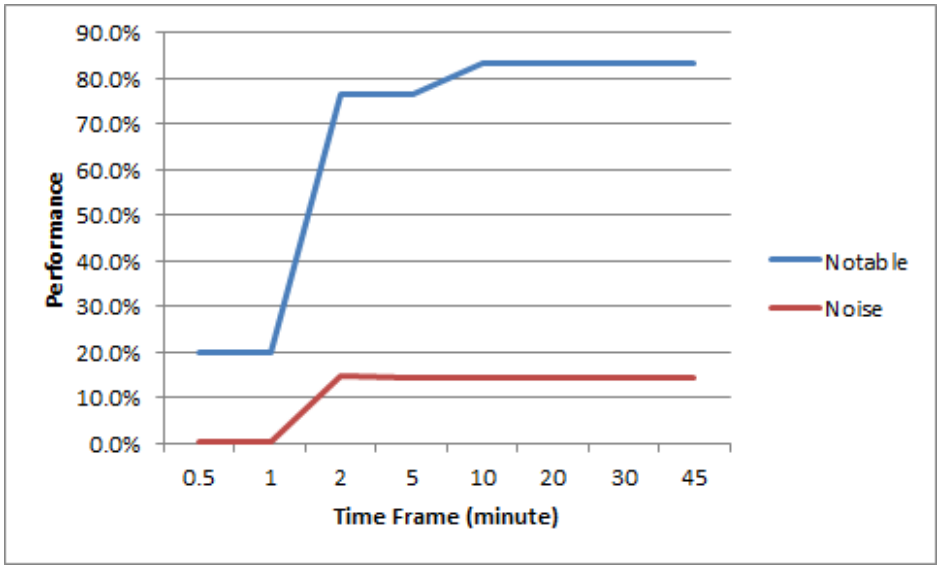

Figure 7.9: Best performance for the Case Private 1 network size 7x7

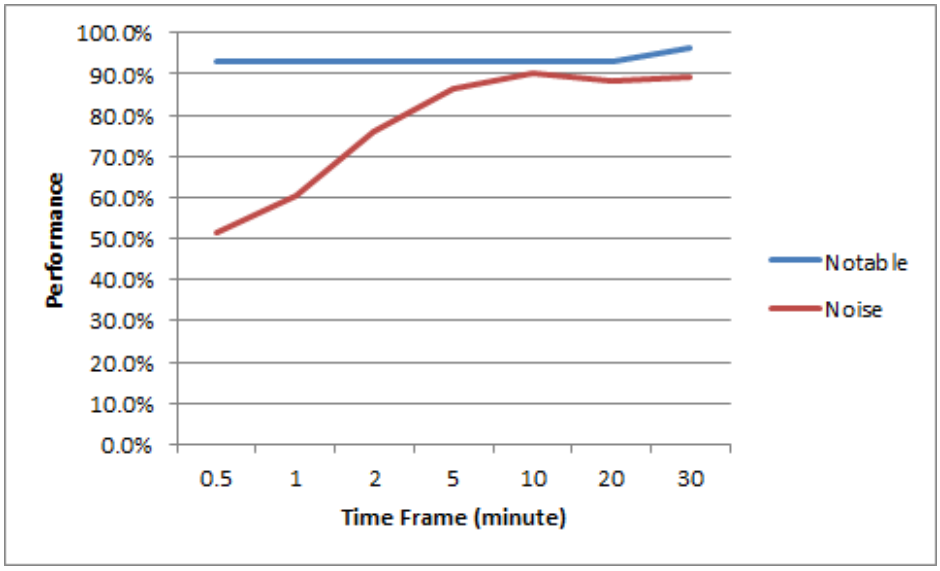

Figure 7.10: Best performance for the case Private 2 - network size 10x10

As illustrated in the above figures, the best time frame length for obtaining a good performance is with a relative small time window. In most of the cases (apart from the case Private 1), the smallest time frame length 0.5 min provides the best notables to noise ratio. However, there are situations in a few cases where the timeframe lengths of 1 or 2 minutes also generated good results.

For the case Public 1, the network sizes 3x3 and 5x5, with a timeframe length of 2 minutes provided the best notables to noise deviation with about 40% (with the proportion of notable being higher). Notice that the values obtained for these network sizes are consistent with the results obtained for the other analysed cases.

As for the case Public 2, the best performance for the algorithm took place within the network size 9x9 along the timeframe 0.5 minute. The notable files percentage was higher with nearly 50% than the noise files. Given the considerable larger amount of files subjected to the analysis than the case Public 1, a larger network size was more suitable to detect large portion of notable files versus the noise files.

Considering the case Private 1 (as demonstrated in Figure 7.9), the best performance was obtained when implementing higher network size value than the cases Public 1 did, but less network size value than the case Public 2. Such a result does support the theory of "larger data requires bigger network size", the input data of the case Private 1 was larger than the case Public 1's, but less than the case Public 2's. Notably, the performance in this case was far different from performance in the other cases. For instance, the percentage of the detected notables sharply increased from 20% to 85% after the timeframe 2 minutes; while the percentage for the noise files remained steady at 14%. This performance pattern remained steady from the second minute until it has reached the maximum given timeframe which was 45 minutes. This fact motivated additional analysis of this particular case to determine the reason for such a result. Taking into the consideration the fact that the analysed data was originated from a portable device, this could be the reason for such a result.

For the case Private 2, the best timeframe corresponding the given network size is 0.5 with network 9x9 & 10x10 with a performance of 42% different between notable and noise files (with the proportion of notable being higher). Due to the large amount of files subjected to the analysis, this was expected when considering the result from the previous cases that contained a smaller number of analysed artefacts.

7.4.7 Result Summary

The Table 7.8 illustrate the detection rate of the AEP for the examined cases and the output of the results that have been achieved. It can be seen that both cases Private 1 and Private 2 have a 100% success rate in detection of notables. Case Public 1 also has a very high detection

success rate at 94%, indicating that the AEP approach yields a very high success rate, competitive with contemporary methods. In contrast, the case Public 2 yielded only an extremely low 8.5% success rate of detection on the notables. This would seem to be a clear sign that the approach would not work for all made up cases. This fact, however, is agreed by the experts that were questioned, whereas if the case had been a real-life one with data from a present case, results would have yielded a much higher success rate.

Case ID	Artefacts		Total Noise	Notables		Detection Rate %
	Total	Analysed		Total	Detected	
Public 1	11,638	2,000	3,372	796	753	94.5
Public 2	22,373	6,179	5178	11,696	1,001	8.5
Private 1	6,543	3,471	3441	30	30	100
Private 2	3,456,219	118,690	118409	281	281	100

Table 7.8: Detection rate of the approach (AEP)

7.5 Conclusion

The analysis of the performance was focused on two main aspects of the AEP: influence of the number of iterations, and influence of the two design parameters (i.e. network size and time frame length). Concerning the number of iterations, it suggests that in most cases two iterations would deliver about the same performance (in terms of percentage of notables being detected) as four iterations were applied, although proportion of noise files increases. This indicates that applying less iteration may be more beneficial in terms of the entire processing time, the detected notables and the notable and noise ratio. Also, using more than one cluster during the process is imperative as the algorithm would continue looking for more potential relevant artefacts that may be contained in the other clusters and may from other SOMs. Such an approach would efficiently be able to insure that the clusters are desirably processed and notable artefacts are detected to a high extent. Concerning the influence of the time frame length, the general trend is that a longer time frame leads to larger percentages of notables and noise files; however, the ratio of the two types decreases with the increase of the time frame length, meaning that as the time frame becomes longer, there are relatively more noise files being selected, than notable files.

When the ratio of the notables to noise files is taken as a measure of performance, it is interesting to look at the recommended choice of the time frame length and the network size

value that provides the maximum ratio of notables to noise files. From network size perspective, the experimental result shows that a smaller number of artefacts within a case with a smaller network size value would create with a better result. For instance, the case Public 1 with the smallest number of artefacts(i.e. 2,090 files), thus new work size 3x3 was the idea for an optimum result. While for the cases with relatively larger number of artefacts, larger network size achieved a better result in terms of ratio of notables to noise. For instance, the cases Public 2 and Private 2, this was the case as the case Public 2 contained a total number of 6,179 files and the case Private 2 had the largest number of artefacts with a total number of 118,690 files. Also, as the case Private 1 contained higher number of artefacts than the case Public 1, but less than the other two cases (Public 2 & Private 2), the best result was achieved when a medium network size was given such as network size value 7x7.

Regarding the influence of the timeframe, the smallest time frame (i.e. 0.5 minute) applied in the experiment provided a good result for three out of four experimented cases (apart from the case Private 1).. For the case Private 1, the performance was very good after the timeframe 2 minutes was applied. As illustrated in Figure 7.9, by applying the timeframes between 2 minutes and 45 minutes, the results in terms of notables were excellent as the notable percentage varied between 77% and 85%. This could be caused by the fact that the suspect could have heavily accesses the notable files within the given timeframes on the portable drive. This case is also with the least number of notable artefacts (but not the least amount of data) , adding another explanation for the peculiar results encountered.

Finally as illustrated in Table 7.8 the performance of the AEP algorithm on three out of four examined cases (Public 1, Private 1, and Private 2) is as accurate as the results would be attained when analysed by a human investigator. The percentage rates for the detected notable files is considerable high for the case Public 1 with a rated of 94.5% of the existed notable files in the case, and for the cases Private 1 and Private 2, all existed notable files in these two cases were detected by the AEP algorithm. However for the case public 2, the percentage of the total notable artefacts to the detected notables is considerably low. The reason behind such a large deviation between the two outputs was due to the fact that a large portion of notable artefacts in this case are belonged to deleted executables. As a result, important and fundamental data such as the associated metadata was missing and not resourced; hence less convincing results. This is a limitation of the proposed approach. Nevertheless, the AEP approach works in 75% of the examined cases as most or all of the

relevant files existed within the analysed artefacts, especially in the real cases. These good results demonstrate that the proposed AEP algorithm can be used to automatically examine cases with little effort from human investigators; as a result huge amount of the time could be saved and more cases can be examined within a given time.

8 A Novel Automated Forensic Examination Architecture

Having established a clear picture about the problem that faces digital forensics, it is imperative to consider a new system that is able to deal with such an issue in a manner that still follows forensic methodologies and best practice. Based upon the promising experimental results from Chapters 6 and 7: the SOM cluster technique has the ability to cluster artefacts within digital forensic images and the AEP is able to automatically capitalize the outcome from the SOM cluster and detect large proportion of notables within a case, this chapter proposes a novel framework that can enable the automating digital forensic analysis process from an operational perspective. A detailed description about the entities of the proposed framework is presented with highlighting the functionalities and the interaction between these entities.

8.1 Introduction

Without substantial level of automation, the process of digital forensics would stand little chance to survive on facing the immense number of computer crime incidents and the growing volumes of data. In addition, triage tools also have certain limitations that need to be addressed and it can be achieved through the process of automation. Also, due to possibility of using of various anti-forensic techniques (such as data hiding and artefact wiping), it is important to understand the technical capability of the suspect; hence invaluable triage time can be allocated accordingly: longer time should be given on the case where the suspect has more IT background (i.e. it is highly likely some level of anti-forensics may be applied). Moreover, as the case may be created under different background (e.g. network hacking, Intellectual Property theft, and child pornography), various legal aspects should be taken into consideration, such as legislations of individual country and rules of multinational corporations. Obviously other challenges (both existing and unforeseeable) should also be considered when dealing a case, such as processing speed, human resource and sophistication of the tools. all these challenges will make the process of analysing a suspect's device more time consuming which in turn could result into many pending cases or/and refutable evidences. It is envisaged that majority of forensic cases will face these challenges to a

certain degree due to the nature of IT development (e.g. large hard drive spaces) and the use of encryption (whether intentionally or not).

To overcome the three main identified challenges (i.e., automation, technical competency, and legislative issues) and based upon the findings of Chapters 6 and 7 (i.e. the use of SOM and AEP to automatically identify notables), a mechanism was researched and developed to enable automating the process of detection and extraction of evidence from forensic images. This approach was one of the promising solutions that could contribute in solving the problem. To achieve the research goal, the proposed mechanism "Automated Forensic Evidence" (AFE) consists of the following:

1. Computer crime Profiling – the development of technique(s) for the automatic detection of forensic artefacts
2. Legislative Requirements – understand what implications legislation can have on an automated tool – particularly in terms of what aspects of the data it is (or not) permitted to access in any given situation
3. Technical Competency Measure – an attribute required to determine the sophistication of the examination required

According to these three key components, they demonstrate that the proposed approach also fulfil the requirement of adaptation. Obviously, the volume of case image data is a key obstacle for any tools in the future; as demonstrated in Chapters 6 and 7, the computer crime profiling technique can deal with real life cases ranging from a few gigabytes to half terabytes. Therefore, it should be able to process larger amount of data as long as the hardware supports. Also, a parallel process could be introduced to deal with massive amount of data, similar to what the bulk_extractor offers. Moreover, the Technical Competency Measure is a key element that deals with the adaptation issue as it automatically offers the level of attention required during an investigation. This ability enables the proposed mechanism to solve case with various backgrounds. Details of the AFE are presented as follows.

8.2 Automated Forensic Examiner

In order to realise the AEP and Technical Competency (TC), it is necessary to design an architecture that could support the algorithm discussed in chapter 7. As illustrated in the

Figure 8.1, the architecture comprises of a number of key processing stages: Forensic Pre-Processing, AEP, TC, Visualizer, Profiler Refiner and Report, and data storage elements.

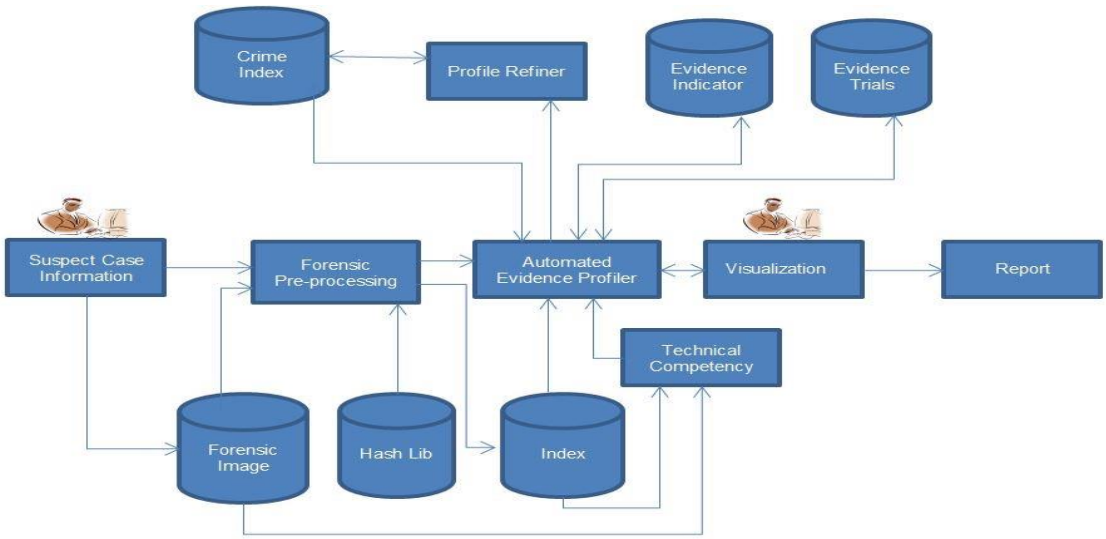

Figure 8.1: AFE Architecture

8.2.1 Suspect Case Information

During 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. Upon receiving the image, the available information about the case (e.g. crime nature, seizure date, image size etc.) is fed into the system. Also, the suspect information such as his job nature, qualifications and IT technical skills are highly considered and provided to the AFE system. This information is fed through a graphical interface that is dedicatedly provided to receive various inputs, including the input features that would vary from case to another according to the case nature. Giving the fact that such information forms the base of the investigation process, it is anticipated that this step would enable the investigator estimate the analysis time required for a case depending on the given information.

8.2.2 Forensic Pre-Processing

This stage will undertake a variety of standardised forensic processes upon the image, including a hash analysis for known and notable files, files signature analysis, extraction of compound files, data and metadata carving, keyword searching (based upon entered suspect information and pre-defined search criteria) and indexing. The primary role of this stage is to

reduce the dataset and effectively sort the "wheat from chaff" in a manner that the relevant information gets separated from unnecessary information. It is worth mentioning that these processes are initially performed by a globally known forensic tool which in this case could be the FTK which output would serve an input to the AFE system. The FTK outputs are exported in CSV format as raw data which is then enumerated to enable the AEP algorithm perform the clustering and iteration processes accordingly.

8.2.3 Indexing

The Indexing component permits parsing of the data so that it gets stored in a manner that makes information retrieval efficient later on. Considering the massive data size that is to be dealt with today, indexing mechanism has a significant impact on the time taken to manage a case. Thus, in the absence of such indexing, it would unnecessarily consume 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 AFE is provided with an ordered and reduced dataset from which to perform its analysis, thus saving time and effort at once. Prior to doing so however, the technical competency process is utilised to appreciate the type and level of analyses required in a case and to provide the examiner with an estimated analysis time accordingly.

8.2.4 Technical Competency

Through an analysis of the complete image (as standard programme files and data might be removed via the hashing process), the Technical Competency component provides a list of advanced analyses that need to be undertaken depending upon the identified criteria. Notably, it will also provide an overall measure for the suspect's technical competency in order to provide the investigator an appreciation of the case complexity on which the required analysis level and time will be estimated based on. Moreover, the suspect's system configuration is considered among the measuring process. The process includes, but not limit to the registry keys setting that could prevent updating last accessed timestamp, overwriting metadata, and any other activities that could prohibit or complicate the digital forensic process. Furthermore, beside the installed applications and the suspect's system configuration, the information about the suspect himself such as his qualification, job nature, computing experience etc. will be fed to the system by the investigator. This information

normally provided by the law enforcement upon handing over the seized devices to the digital investigators team. Together, those multi resources form considerably consistent pattern to measure the technical competency level of the suspect.

8.2.5 The Automated Evidence Profiler

As previously illustrated, 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 Trails occurs. The different types of data including but not limited to graphics, text, audio, timestamps, contacts, email communications, browser behaviour are mapped and updated to make a profile of the information within the case.

The AEP respectably plays the major role among the other components within the AFE through the multi-functions and process that are performed. This component is fed with the required information as basic input from various components (i.e. Forensic Pre-processing, the NSRL index database, Technical Competency etc). Subsequently, upon processing the input artefacts, the AEP interacts with the other AFE components mainly Evidence Trial, Evidence Indicator, Profile Refiner, Crime Index etc. Following to the process, the AEP interacts with Visualization component in two different directions for two purposes, updating the Visualizer and in return receiving the feedback. Furthermore, advanced analyses will also be undertaken depending upon the outcome of the Technical Competency analysis. Obviously, the above stated practice of the AEP indicates that it performs the required analysis in a case-management approach along a process starts from receiving input in a form of pre-processed artefacts to providing the Visualizer with filtered and refined data.

8.2.6 The Crime Index Database

It is the component where the indexed database that 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. Upon receiving the refined profile from the refining entity, the Crime Index database is updated and passed to the AEP for processing upon request. Through the Profile Refiner, this component permits the system to adapt to the changing computer crime environment as new terminology and artefacts are created. Whilst, there are several data sources within the AFE system that the AEP utilises, the Crime Index Database acts as a key catering source. With

the presence of Crime Index Database, the time taken for case analysis in future would be relatively shorter than it was expected to be. Such an outcome is logically expected as a result of deploying an equally approach.

8.2.7 The Profile Refiner

This component permits the system to adapt to the changing computer crime environment as new terminology and artefacts are created. It will also update the Crime Index Database with the related information. It is anticipated that once many cases are received, there may be pattern of similarities of the evidence between cases. For instance, artefacts that are found within blackmail cases would be similar to each other. Similarly, artefacts that are presented within child pornography incidents will mainly vary between images and videos in all cases of the same nature. Therefore, the role of the profile refiner is to understand such patterns and to pick up those artefacts whenever they exist. Such a process will eventually result in speeding up the investigation process and reduce both the efforts required by the investigator and the time taken when analysing a case.

8.2.8 The Evidence Indicator Database

The Evidence Indicator Database stores the extracted artefacts that the AEP process has identified; thus presenting a centralised collection of evidence pertaining to the case. The evidence indicator process is basically about mapping the crime against the known document types and keyword list. This is refined both by using an automated process in the AEP as well as based on human examination and associative mapping by the investigators wherein the person looks for typical file signatures based on some concept of a target object in mind (Carrier &Spafford, 2004).

8.2.9 The Evidence Trails Database

The Evidence Trails database is utilised to store the metadata associated to the extracted artefacts whereas the actual artefacts are stored in the Evidence Database Indicator. The evidence trail is more concerned with finding links between different artefacts in order to create a bigger picture by trying to understand the manner in which the artefacts were used. This also helps to find a clue for the next artefact which could be possibly related to the case. Hence, this database is updated accordingly and the AEP tries to differentiate the criminal evidence trails from the normal patterns of system usage. This is an iterative process which is done as far as the automated intelligent system can manage. Any changes made to the

Evidence Trails by other components are fed into the Evidence Trail Database whether it is the manual change or priorities or refinement of the evidence trails. Storing the evidence trails in a separate database also gives an opportunity to keep them in an organized manner from where they can be retrieved, updated or edited as required for use in the AEP and other components of the AFE system.

8.2.10 The Visualizer

This component represents the link between the AEP and the final report output. Recognising that the AEP process will inevitably identify false Evidence Trails and thus artefacts, this process exists to conveniently and usably present the Evidence Trails so that an investigator can discount or decrease/increase the priority of the Trails. Any modifications made to the priorities at the visualizer by the human operator are reflected back to the AEP which in turn updates the relevant databases such as the evidence indicator and the Evidence Trails database. Once the human digital forensic examiner finalizes the priorities of the automated and any changes made therein, the visualizer passes on the information to the next component which is used to generate the final report of the entire process. Hence this stage represents the last lap of the iterative process which is used to continuously improvise upon the artefacts until a relevant evidence trail is generated.

8.2.11 Reporting

This is the final output of the AFE system, and represents the analysis and the results of the entire investigation exercise. The report generation component receives its input from the final output of the visualizer. The AFE would generate the report in such a manner that is readable and could be understood by non-technical individuals included those serving in the judicial firms. Thus, generated reports would be used in the legal context for prosecution purposes and help the investigative agencies to pursue the matter in the courts of law after the technical analysis part is completed. It culminates the concept of cyberprofiling based on technical competency parameters and translates the raw data into a collection of relevant artefacts and the associated evidence trails in a certain order of priority. The overall purpose of the entire exercise is to significantly reduce the processing time for extracting this information from the initial raw data which will help to speed up the investigation process. The AFE may not be in a position to totally eliminate the need for human intervention but it is certainly necessary given the explosion in data volumes and rising number of digital crimes in the recent past. The reporting process however, is not the final stage of investigation but

provides a platform from which the human investigators take on the further course of actions. It is estimated that once the AFE is functional and used in computer crime investigations it would be a significant step in the world of computer crimes investigation.

8.3 Technical Competency Concept

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 according to applications and settings presented on a system. 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 can be utilised for. Modifying or changing basic configuration options such as the cluster size would also provide intelligence that the suspect has been modifying settings, possible to the advantage of hiding data. Therefore, the higher the technical competency of the suspect, the deeper the analysis should be provided during the investigation. Table 8.1 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. Also, this table is utilised mainly for illustration purpose as both the criteria and their associated impact required further analysis in order for it to be used in a real system.

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
Wiping software	High

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

Table 8.1: Technical Competency Criteria

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. In comparison, if this measure indicated that the suspect was naïve or an ordinary user, more advanced analyses would not be invoked within the AFE and only evidence found during the normal analysis would be passed on for processing.

8.4 Discussion

Despite the fact that previous attempts of overcoming the problem of dealing with large amount of data through triage and partial automation have enhanced the digital forensic field, the need for a more comprehensive automation system is vital to meet the future requirements of the domain. Criminal Profiling is one of the 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 to assist analyzing future cases, thus reducing the time taken to prove or otherwise the case.

The proposed AEP features an iterative-based approach to identify potential evidence and performs associative mappings to related events, enabling the AFE system to create evidence trails that is able to filter and refine the processes. The evidence trails are created to provide an artefact 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 trails surrounding the use of that artefact, with the intention of locating further artefacts (whether they are images or information pertaining to other offenders or what the suspect used them for).

In order to undertake the analysis, AI techniques such as the SOM clustering was 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 Trails.

8.5 Conclusion

The proposed approach 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 reduce the dependency on the human investigators, thus freeing up valuable expertise to investigate more complex cases. Therefore, it is critical that the proposed approach is empirically evaluated in the future. With this in mind, an expert based evaluation on the proposed approach is presented in the forthcoming chapter.

9 The Evaluation

The chapter begins with a description of evaluation methods in general and the justification of the selected method. Further sections of the chapter describe the evaluation scope, and illustrate the implemented approach. A brief summary of the PhD research and its key results was prepared in the form of a presentation, which was used for briefing potential participants, informing them about the key contribution of the research and also taking their consent for the evaluation. The participants, who were all experts within the domain of information security (some were specialised in digital forensics), were carefully selected; they were interviewed via a set of rigorously designed questions, permitting them to provide feedback in a constructed and focused manner. This was followed by a detailed interview and discussion with the experts on the various aspects of the proposed study. The chapter concludes by discussing the findings of the evaluation.

9.1 Introduction

One of the main research areas in this thesis has focused on the development of an Automated Forensics Examiner (AFE) which aims to ease digital forensic examiners' daily workload by trying to speed-up the analysis process with artificial intelligence techniques, enabling relevant files to be quickly identified within a given time. Hence, it is critical to assess the actual effectiveness and possible implementation of the proposed AFE in the area of digital forensics, and also it is equally important to evaluate the proposed approach and the associated development work to fulfil the research requirements. As a result, the evaluation of the research outcomes was necessary, and appropriate methods were selected to ensure the most informed results can be obtained from the evaluation stage.

9.2 The Evaluation Method

The purpose of research is essentially to explore new findings that can make a valuable contribution to knowledge. The traditional or conventional approach to research focused on producing general laws which accommodate the given facts (Pant, 2015). Thus, selecting an appropriate evaluation methodology is a key factor in measuring the outcome of the conducted research. New approaches to research have developed which are different from the concept of conventional research. One such technique which has challenged the conventional

approach is participative research. Participative research involves the participation of people (usually stakeholders), who should be able to provide the right answers and feedback to the questions in place; hence, participants can act as evaluators in their own right.

There are three main steps in conducting participative research, namely: identification, sampling and gaining information (Krishnaswamy, 2004). The first step involves identifying the people who can give the required information, as not everyone is suitable for a given evaluation. The second stage is to find a diverse (yet small and manageable) sample of people which represent different stakeholders on the issue. The final phase is the planning to gain reflective feedback and opinions from the selected participants in a convenient manner.

After careful analysis of the advantages and disadvantages of the conventional research and participative research methods and taking other aspects (such as time and budget constraints) into consideration, it was decided that the participative research was to be utilised as the evaluation method. Obviously, it would be the best option to have a face to face interview with participants. However, as many identified experts were located in various geographical areas and even in different countries, it was deemed that the best way was to use video conferencing as it is one of the most convenient and popular methods for meetings.

As mentioned previously, stakeholders would be the best candidates to perform the evaluation. There are two main stakeholders within the domain of computer forensics: practitioners and researchers. The former mainly focuses upon the functionality side of a forensic tool, while the latter considers the contribution to knowledge as the primary goal of research work. Due to the nature of the proposed work (mainly research based) and the current stage of research (primarily at the conceptual phase), it seemed logical that researchers were the best option for the evaluation process. Indeed, researchers and academicians are better able to foresee problems and realise how the demand for the required development will fulfil future requirements, while the practitioners and law enforcement personnel are more likely to concentrate on the current situation. Based upon their knowledge, expertise and experience within the domain of information security (mainly computer forensics), 23 academics were targeted as the potential expert evaluators.

9.3 Evaluation Scope

Existing digital forensic tools are created for assisting the investigation at various stages, such as data collection, triage, examination and analysis. As a result, it is necessary to evaluate the usefulness of the approach developed in this research in order to find out to what extent the developed approach can contribute to the computer crime investigation in addition to existing tools.

All 23 identified academic experts were contacted by email regarding the research evaluation. They were informed that the purpose of the evaluation exercise was purely to determine the effectiveness of the study in terms of its application in the digital forensics domain. For those who agreed to participate in the evaluation process, a consent form was sought from them.

The evaluation process was divided into two main phases. Firstly, a narrated Microsoft PowerPoint presentation was prepared which explained the approach in detail, including its scope, aim and the suggested solution. The PowerPoint presentation was then sent to all the participants through email, allowing them to view the file at their convenience.

The second phase of evaluation consisted of contacting each of the participants individually and conducting an interview by asking a pre-defined set of opened ended questions which will be outlined in the next section. The interviews were conducted via Skype at the convenience of the participants as they were all engaged in their full time professions with associated commitments. The following sections give an overview of the actual interview questions, the background of the experts and their feedback.

9.4 Evaluation Questions

Based on the requirements of the evaluation, a total of 13 questions were designed for the evaluation task and they are listed as follows:

1. Do you agree that the taken research problem is valid?

2. What do you think about the approach? Do you think it is scientific?

3. Do you think that it is important to have a digital forensics tool that is able to consider the legislative aspects?
4. What do you think about SOM's capabilities in terms of digital forensics?
5. Do you think it is important to find the link between the artefacts?
6. Do you think the evidence trial indicator would help determine if several parties were involved in the investigated case?
7. Do you think having a crime index data base would help in speeding up the analysis process in future?
8. What do you think about measuring the technical competency of the suspect? Is it important to do so?
9. What do you think about implementing timeline analysis approach?
10. What do you think about the attained experiment results? Is it satisfying?
11. From the result illustrated in the previously provided PowerPoint presentation, do you think the research has a sufficient contribution?
12. How realisable/attainable/possible do you feel this system is?
13. What do you feel are the particular strengths & weaknesses of the approach?

As demonstrated above, these questions were mainly created to evaluate the entire approach towards development of the AFE and the various perspectives that were considered while designing the approach. Questions were asked about the efficacy and usefulness of SOM clustering techniques and their relevance for solving digital forensic problems. Questions were also designed for the technical competence measuring capabilities of the tool and whether it had any significance in real world situations. Then, the participants' opinion towards the aspects of timeline analysis, linking artefacts and the capabilities of the other components within the AFE were sought. Beside the technical aspects, another important factor, the legislative aspect of a tool which technical experts usually tended to overlook was mentioned in the terms of the AFE.

The key experimental results from Chapters 6 and 7 were also presented to the participants, providing an opportunity to get their opinions and feedback on the effectiveness of the approach and the experimentation as a whole. Furthermore it requires careful afterthought and a detailed study by the participants before they can actually comment on the questions in depth. This ensures that the expert considers the research from all aspects and then gives their opinions, which can also be used to evaluate the entire exercise and further improvement in future studies.

9.5 The Participants

This section describes the selected participants for the purposes of evaluation and the justification for choosing these participants. Since the research is related to digital forensics, the chosen experts were taken from appropriate backgrounds so that they had the necessary knowledge and expertise to evaluate the research and the AFE. A list of candidates was made and the final six participants selected are listed below. Aside from the academic and professional backgrounds, the selection criteria were also based on the willingness of the candidates to be interviewed and their availability. These six participants produced enough insightful information to be sufficient for the purposes of evaluation and so no further candidates were chosen.

9.5.1 Dr Robert Hegarty – Manchester Metropolitan University – UK

Dr Robert Hegarty, Senior Lecturer at Manchester Metropolitan University in Manchester (United Kingdom). Email: r.hegarty@mmu.ac.uk – Skype interviewed on 27 July 2015

Dr Robert Hegarty is a senior lecturer in computer security and digital forensics from the School of Computing, Mathematics and Digital Technology at Manchester Metropolitan University (UK). Also, as a member of the Future Networks and Distributed Systems (FUNDS) research group, he carries out research in the areas of complex distributed systems and associated issues including but not limited to, network security and computer forensics. As a result, Dr Hegarty has published several papers in the aforementioned fields and presented related research outcomes in various international conferences. Dr Hegarty's previous experience also includes being a part of the PROTECT Research Centre at Liverpool John Moores University, UK, where he worked on problems related to computer

and network security. He has also been involved in imparting training related to digital forensics to various law enforcement officers. According to his excellent academic background in the field of digital forensics, Dr Hegarty became an ideal candidate to evaluate the Automated Forensic Examiner. He was interviewed on 27 July 2015 via Skype.

9.5.2 Dr Paul Sant – Bedford – UK

Dr Paul Sant, Associate Dean (Quality & Management) at University Campus Milton Keynes in the University of Bedfordshire (United Kingdom). Tel: +44 (0)1908 295809 – Skype interviewed on 29 July 2015

Dr Paul Sant is the Associate Dean (Quality & Management) in the Department of Computer Science and Technology at University Campus Milton Keynes in University of Bedfordshire (UK). His research interests include computer security, forensics, security modelling in pervasive environments and trust modelling. Dr Sant is a journal reviewer in several reputed journals and is an active researcher in the aforementioned areas, especially digital forensics. He is also a member of the British Computer Society and associated with the European ECENTER Project. Dr Sant's areas of interest combined with his academic credentials certainly make him a suitable candidate who can provide value feedback in the evaluation of the proposed Automated Forensic Examiner. Using Skype, Dr Sant was interviewed on 29 July 2015.

9.5.3 Dr Christos Kalloniatis – Aegean – Greece

Dr Christos Kalloniatis, Department of Cultural Technology and Communication, University of the Aegean, Email:chkallon@aegean.gr - Skype interviewed on 05August 2015

Dr Christos Kalloniatis holds a PhD from the Department of Cultural Technology and Communication of the University of the Aegean and a master degree on Computer Science from the University of Essex, UK. Currently he is an assistant professor in the Department of Cultural Technology and Communication of the University of the Aegean. He has alo served as a visiting professor in many European Institutions. Dr Kalloniatis' main research interests are the elicitation, analysis and modelling of security and privacy requirements in traditional and cloud-based systems, privacy enhancing technologies and the design of information system security and privacy in cultural informatics. He has authored several refereed papers

which were published in international scientific journals and conferences. Prior to his academic career Dr Kalloniatis worked at various places within the Greek public sector including the North Aegean Region and Ministry of Interior, Decentralisation and e-Governance. He is a lead-member of the Cultural Informatics research group as well as the privacy requirements research group in the Department of Cultural Technology and Communication of the University of the Aegean; also, he has a close collaboration with the Laboratory of Information & Communication Systems Security of the University of the Aegean. Dr Kalloniatis has served as a member of various development and research projects. Based upon his knowledge and experience in the domain of information security, Dr Kalloniatis was contacted with regard to the evaluation of the proposed Automated Forensic Examiner. Dr Kalloniatis was skyped on 05 August 2015.

9.5.4 Dr John Haggerty – Nottingham Trent University – UK

Dr John Haggerty, Senior Lecturer in the School of Science and Technology at the University of Nottingham (United Kingdom). Email: john.haggerty@ntu.ac.uk – Skype interviewed on 01 Septemper 2015

Dr John Haggerty is a senior lecturer in the School of Science and Technology at the University of Nottingham (UK). His teaching areas include network security, information security, internet technologies and digital forensics. He is an active researcher as he regularly presents his research work at various conferences. Recently, he contributed to a new security tool named XDet, which can detect malicious data in cloud environments. According to his experience in the digital forensics arena and in the creation of security tools, Dr Haggerty is an ideal individual who can provide a thorough evaluation of the AFE tool, including suggesting any potential shortcomings and drawbacks that can be used as a basis for further improvement. Dr Haggerty kindly offered his time on the evaluation on 1 September 2015.

9.5.5 Professor Andy Jones –Professor at Edith Cowan University in Perth, Australia

After working for 25 years in the British Army Dr Andy Jones became a manager as well as a researcher and analyst in the area of Information Warfare and computer crime at a defence research establishment. In 2002, Dr Jones left the defence environment and became a principal lecturer at the University of Glamorgan (now the University of South Wales), lecturing on the subjects of Network Security and Computer Crime and researching on the

topic of threats to information systems and computer forensics. Within the same institution, he also developed and managed a well-equipped computer forensics laboratory, and took the lead on a large number of computer investigations and data recovery tasks.

In 2005 Dr Jones joined the Security Research Centre at BT where he became the Chief Researcher and the head of information security research. During his time at BT he managed a number of research projects and led a series of projects into residual data on second hand media. In 2009 Dr Jones became the Programme Chair for the Information Security (Masters level) course at Khalifa University of Science, Technology and Research in Abu Dhabi in the UAE. Currently Dr Jones holds visiting professorship from Edith Cowan University in Perth (Australia), the University of South Australia in Adelaide (Australia), De Montford University (UK) and the University of South Wales (UK). Prof. Jones also holds a PhD in the area of threats to information systems. He has authored seven books on topics including Information Warfare, Risk management and Digital Forensics and Cyber Crime, and has also published more than 100 papers on the aforementioned subjects. Dr Andy Jones was interviewed on 17 September 2015.

9.5.6 Dr Theodore Tryfonas – Bristol University

Dr Theo Tryfonas, Senior Lecturer in Systems Engineering at the University of Bristol (United Kingdom). Email: theo.tryfonas@bristol.ac.uk– Skype interviewed on 02 October 2015

Dr Theo Tryfonas is a senior lecturer in systems engineering at the University of Bristol (UK). Dr Tryfonas is deeply interested in cybersecurity and computer crimes and has a wide range of research interests especially in the domain of digital forensics. Regarding digital forensics, Dr Tryfonas is both a researcher and a practitioner. He has authored many research papers in the digital forensic domain and also presented evidence in court cases that were related to digital forensic crimes. Given his first-hand experience not only of the digital forensics technologies but also having exposure on the legal front, he was identified as one of the most suited researchers to discuss the legislative aspects of digital forensics, an important part of the proposed AFE.

9.6 The Evaluators' Feedback

In this section the answers for the questions that were posed to the experts during the interview process are presented and discussed in detail, with the answers being presented in a question by question manner. In this way, comparisons of expert's feedback can be easily carried out, allowing for a more comprehensive method of evaluation to be obtained.

9.6.1 Validity of the Research Problem

The research question was seen as valid by all the participating experts. From the presentation, Dr Christos Kalloniatis found the research topic interesting and the statistics gap was precisely covered. Also, Dr Robert Hegarty and Dr Paul Sant agreed that the subject is a massive research problem. Given the fact that computer crime has and will continue to increase, more effort should be given to enhancing the speed of the artefact identification (preferably via automation) to ease the workload of digital forensic investigators; this background was agreed by Dr Paul Sant, Dr John Haggerty and Professor Andy Jones. Also, due to the changing nature of computer crimes and the increase in data volumes, Dr John Haggerty, Professor Andy Jones and Dr Theo Tryfonas all agreed that automating the process is one of the key approaches for solving the research problem.

9.6.2 Efficiency of the Suggested Approach

After thoroughly examining the proposed framework and technology of the research, all of the evaluators agreed that the chosen approach is scientifically sound and it precisely addresses the research problem. Despite SOM being an adequate method to solve the research problem, Dr Paul Sant, Dr Cristos Kalloniatis and Dr John Haggerty highlighted that additional justifications on why SOM was chosen are required. Also, the research would have benefitted greatly if more experiments had been conducted using other clustering techniques. Having perfectly understood the proposed approach, both Dr Andy Jones and Dr Theo Tryfonas suggested that the quality of the presentation could be enhanced with more details and additional information about the proposed method being presented.

9.6.3 Importance of the Legislative Aspects to the Forensic Tools?

In response to this question, the majority of the evaluators were in clear agreement: a tool that could be used to consider the legal aspects of forensic examination based on certain criteria would be ideal. Without a tool considering legislative aspects, it is difficult for both

practitioners and law enforcers to cope with such a problem, and thus there was an important need for a tool that is able to validate the investigation from legal aspects (Dr Robert Hegarty). A tool that was able to incorporate both legal and technical aspects would bring much to the field of study (Dr Christos Kalloniatis). It would however be quite a challenge given the different related laws in various locations such as the UK, Europe and the Middle East (Dr Paul Sant). Indeed, without having a tool that fulfils the legislative aspects, it may make the investigator get caught in legal implications that may result in evidence being refuted (Prof. Andy Jones). In contrast, Dr John Haggarty had a different view as existing tools such as e-Discovery had already claimed to have built-in legislative aspects. Hence attention should be given to address newer problems, such as that the current digital forensic tools do not currently consider legislative aspects. Despite Dr Theo Tryfonas agreeing that it is important to have a digital forensics tool with built-in legislative aspects, he also suggested that it would be difficult to have such a successful tool in reality due to legal contexts where different legislations are in place.

9.6.4 SOMs Capabilities in Terms of Digital Forensics

All the experts considered that the use of SOM is a novel and interesting solution to the problem of automatically identifying artefacts relevant to the case. Also, the initial experimental results look promising and the result could be enhanced with additional experimental studies (Dr John Haggerty; Dr Andy Jones). Moreover, the strength of the research can be improved by testing the idea using other artificial intelligent method such as neural networks (Dr Robert Hegarty) and other clustering tools (Dr Christos Kalloniatis). Furthermore, the approach could be better understood if additional information were offered (e.g. chapters from the thesis) (Dr John Haggerty). In contrast, after noticing that only 8% of the notable files were detected by the proposed method for the case Public 2, Dr Andy Jones pointed out one of the limitations of the proposed method was the dependence on timestamps and should therefore be considered for future research.

9.6.5 Importance of Linking Between the Artefacts

The majority of responses agreed that finding the link between artefacts while investigating a computer crime is very important. By linking the artefacts together, a more accurate view can be obtained, explaining the key questions (i.e. what, why, how, when, who and where) of what digital forensics tries to answer. Nevertheless, the use of virtualisation and cloud based storage can make the situation more challenging as full copies of files may not be available

depending on how they have been used (Dr Paul Sant). Also, as the technology constantly changes and develops, the ability of the proposed method to handle different categories of files requires further validation (Dr Christos Kalloniatis). In comparison with other experts' views, Dr John Haggerty believes the importance of finding the link between artefacts depends on the nature of the case and finding the link may not be required for all cases. Also, the need to find the links for more supporting evidence and using the links to link evidence to multiple sources are two different things. For instance, links between artefacts are important when dealing with high-tech crimes by intelligence services, while it may not be applicable for those cases handled by the ordinary police force.

9.6.6 Usefulness of Evidence Trial Indicator

All the experts were convinced that the Evidence Trial Indicator is a useful and important function of the proposed AFE. Indeed, it is widely expected that more than one suspect is often involved with an open case; hence by using the evidence trial to find additional information based upon their time stamp metadata, other involving parties can also be identified and eventually brought to justice. Dr Haggerty and Dr Jones also suggested the use of the Evidence Trial Indicator can be based upon a number of factors such as the nature of the case, whether the investigator needs to ascertain whether or not multiple parties were involved, and the available information for the case. Moreover, Dr Haggerty pointed out that by using the Evidence Trail Indicator, additional work will be generated for the digital forensic investigator as more information from other potential parties needs to be analysed, and this will increase the time taken to analyse a case. However, this pressure can be eased if the analysis is carried out automatically, enabling the case to be examined thoroughly and bringing more criminals to justice.

9.6.7 Usefulness of Crime Index Database

The experts' views on the usefulness of a crime index database are equally divided: while Dr Robert Hegarty, Dr Paul Sant and Dr Cristos Kalloniatis agreed, the other three experts had a different view. From the supporting side, evaluators considered the background information from previous cases is useful as it can be used to discover, consider or evaluate similar cases, and what the potential outputs and outcomes of work might be, as well as those who may potentially come back to reoffend. As a result, a more informed and accurate analysis can be carried out. Also, the UK police force could benefit from a centralised crime information database as no such repository exists at the moment (Dr Rober Hegarty). In comparison,

Dr John Haggerty thought that the reality of such a database's usefulness was uncertain due to the lack of information provided on the evaluation slides. Also, Dr Andy Jones and Dr Theo Tryfonas thought that the usefulness of the crime index database depends on the data that was stored. For example if the crime index database had certain information linking to crimes of a certain nature, the workload of the investigator and the time taken to process a similar case would be reduced. Also, from a long term view, the digital forensic community will benefit from the crime index database as the number of records will increase, hence its information will grow in usefulness.

9.6.8 Measuring the Technical Competency of the Suspect

The majority of the experts thought that it is definitely important to measure the technical competency of a suspect as it would have a great influence on the investigation process. It was important for the investigator to know the technical background of the suspect so the case can be dealt accordingly. For example, if the suspect had a greater competency level than the investigator, the possibility of third party help would be considered (Dr Paul Sant). In terms of how the technical competency of the suspect could be measured, Dr Robert Hegarty suggested that the use of the Operating System with the correlation of other aspects (such as the patching level) could be used. In contrast, Dr Andy Jones and Dr Theo Tryfonas considered that the case background can be used to measure the technical competency level of the suspect. For example, the suspect's technical competency levels would be high, medium and low for hacking, fraud and paedophile cases respectively. In addition, Dr John Haggerty could understand the concept of measuring competency of a suspect. However, he struggled to understand how it might be measured since the presence of hacking or encryption tools in someone's computer may not correctly indicate the owner's competency level since sometimes these tools could be downloaded for curiosity.

9.6.9 Implementation of Timeline Analysis Approach

In response to this question, all the experts thought it is important for the timeline analysis approach to be implemented as the technique is very useful for case analysis. Dr Robert Hegarty stated that the skill level of the developer and time available for the implementation would have an effect on the approach. Dr Paul Sant agreed that it would be important to create a reconstruction and analyse whether it occurred over a sustained period of time in smaller, incremental steps, or whether it was something that occurred with a much larger bandwidth. Moreover, the timeline analysis can be used to aid profiling, providing an idea

about the suspect's interests, browsing activities, and the amount of time accessing their computer (Dr Andy Jones). Finally, Dr Theo Tryfonas endorsed the use of visualisation for the timeline analysis as information about related evidence and the number of people that interacted over cyberspace can be easily interpreted. Also, currently it may be difficult to include this in legal proceedings since the constructed timeline must provide very elaborate visualisation; essentially, the more complex the evidence becomes, the more complex the processing will be. There are many concepts that need to be explained in ways with more clarity that would make more sense to juries and judges in court.

9.6.10 About the Attained Experiment Results

Regarding the quality of the obtained experimental results, all the experts thought the results were interesting and promising. Notably, Dr Theo Tryfonas stated that the positive results demonstrate that the time taken during the investigation process could be improved significantly when the proposed method applied. Also Dr Christos Kalloniatis thought the results clearly showed the capabilities of the AFE. Suggestions on how the results could be consolidated and improved upon were made by the experts, including further explanation about the results (Dr Robert Hegarty) and the algorithm (Dr Christos Kalloniatis; Dr Andy Jones), the use of additional real-life cases (Dr Paul Sant), and validation of the results via other methods (Dr John Haggerty). Also, Dr Andy Jones pointed out that the result of the Public 2 case raised some concerns due to its nature (i.e. a synthesised case) although this may not have been so much of a problem if the method were applied on real cases (as demonstrated by the Private cases).

9.6.11 The Research Contribution

When asked about this question, all experts answered that according to the information presented within the evaluation PowerPoint slides, this research made valuable contribution to the domain of digital forensic investigation as proposed (i.e. by using SOM to cluster artefacts from computing devices) and can be used to potentially save the investigator's invaluable time during the case analysis phase. Indeed, Dr Paul Sant boosted the claimed results by agreeing that the research provides a novel approach in the field which is different to what has been achieved before within the digital forensic domain, adding valuable contribution to the academic research field. In addition, Dr Theo Tryfonas said that by conducting a large set of experiments in a logical manner and obtaining possible results, the proposed architecture that utilised SOM could be proven to work as designed. Experts also

made a number of useful suggestions on how the research can be better understood and improved. Dr Robert Hegarty suggested that further explanations of results within the evaluation slide could be more helpful. Dr Cristos Kalloniatis pointed out that the result needs to be published before the digital forensic field can benefit and the methodology can be strengthened by using extra classification approaches in addition to SOM.

9.6.12 Possibile Implementation of the AFE

Overall, the experts' feedback to this question was very positive, indicating the possibility of implementing the AFE. Initially, the AFE is a good proof of concept as it clearly has a very scientific background, as well as scientific modules (Dr Paul Sant). Also, a large proportion of the system is also considered to be developed (Dr Robert Hegarty). In addition, the user interface that has been built around the AFE architecture is familiar to digital forensic investigators (Dr Paul Sant). Moreover, law enforcements in particular are in need of such a tool due to the overload of work, and it could be used by non-experts, allowing cases to be focused on according to their nature (Dr Andy Jones). In contrast, Dr Theo Tryfonas thought that an automated system would not be able to make the same kind of determination as a human being when it comes to legal matters. For example, applying the proposed system in a civil or corporate case might well be possible but it would not be admissible from a legal perspective. Dr John Haggarty thought that the approach required further research to be able to be functionally implemented.

9.6.13 Strengths and Weaknesses of the Approach

In answer to this question, all the experts provided their views on both the strength and the weakness of the approach. In terms of its strength, the approach is a great attempt to fill the gap within the digital forensic domain (Dr Cristos Kalloniatis); it utilises the artificial intelligent technique (i.e. SOM) to automatically identify relevant artefacts of digital forensic investigation cases (Dr Robert Hegarty), offering a greater speed than the current situation where the investigator must manually sort through files (Dr Andy Jones). Also, the approach attempted to bring things together, through implementing the clustering with timeline analysis (Dr John Haggerty); such an approach will be increasingly important in the future when designing a digital forensic tool (Dr Theo Tryfonas). Moreover, the architecture is clearly and maturely designed (Dr Cristos Kalloniatis), with the use of several novel components, such as the crime index database, the artefact linkage, and visualisation aspects (Dr Andy Jones). Furthermore, another strength of the proposal is its good probabilistic

background, given the fact that forensic cases are not always determinant (Dr Paul Sant). In the meantime, a number of suggestions were also made by the experts, aiming to overcome some of the weaknesses of this research. Initially, additional elaboration on the results would have provided a better view for the experts during the evaluation stage (Dr Dr Robert Hegarty). Also, further experiments should be carried out to consolidate the result outcome, such as refinements of the SOM parameters (Dr Paul Sant), the use of additional cases (Dr Cristos Kalloniatis and Dr Andy Jones), and the use of other clustering techniques (Dr Cristos Kalloniatis). Moreover, the system should be fully developed (Dr Paul Sant) and integrated with existing case management tools (Dr Theo Tryfonas); in this way, digital forensic investigators can take full advantage of this research. Furthermore, additional publications from this project will offer more support for the research side of the digital forensic domain (Dr Cristos Kalloniatis).

9.7 Suggested Enhancements

The feedback given by the expert participants in the evaluation phase led to several suggestions for the enhancement of the research approach as well as development of the Automated Forensic Examiner (AFE). Some of the important points suggested by the experts are summarised in this section.

9.7.1 Validity of Research, Approach & Usefulness of SOM Technique

The majority of the experts agreed that the research was valid and had practical usability in the area of digital forensics. The same opinion was also received regarding the use of the SOM clustering technique. However some of the experts suggested that the study should be extended further with alternative clustering or AI methods and those results should be compared against SOM in the search for a better alternative. Also, while using a single methodology is sufficient, it would be an improvement if additional justification were provided on the usage of SOM (Dr Kalloniatis, Dr Haggerty, Dr Sant and Dr Jones).

9.7.2 Legislative Compliance of AFE

Most of the experts agreed that it is important for the AFE to comply with the legislations as otherwise a practical application of the tool could be challenged from a legal perspective. However Dr Haggerty held the contrary view as tools such as e-Discovery claim that they follow legislative requirements; as a result, this tool would not add much value unless it

addressed some new aspects from the legislative point of view. This is certainly an important insight since any research should further enhance the already existing knowledge and address issues which have not been discussed in previous studies.

9.7.3 Crime Index Database & Linking of Artefacts

The idea of a crime index database was perceived by the participants as novel, except by Dr Haggerty and Dr Jones who were not fully convinced about the usefulness of this component in real world situations. They suggested that its usefulness would depend on the manner in which the crime index database is implemented and the type of data that is stored in the database. In comparison, all experts agreed on the concept of linking the artefacts together to build a clearer picture that can be used to better understand the crime, in conjunction with the timeline analysis approach.

9.7.4 Evidence Trail Indicator and Technical Competency of Suspects

The experts' opinions were united on the usefulness of the evidence trail indicator as it would clearly help to identify potential co-suspects in the case. In addition, Dr Haggerty and Dr Jones pointed out that despite being practical, this needs to be used with care as otherwise valuable time and resources could be wasted. Similarly, the idea of gauging the technical capability of a suspect was important, although some of the experts (e.g. Dr Haggerty) visualised difficulty in actually implementing it and finding a benchmark against which to measure technical competency.

9.7.5 Experimental Results & Analysis

The experts were also convinced about the validity of the experimental results although most of them stated that the results should be explained in more detail, both in terms of their algorithms and the techniques used for analysis; while they understood that only limited information can be presented in the evaluation PowerPoint slides. Dr Jones expressed concerns over case 2 but since it was a fictional case, it possibly represented the worst case scenario and the situation would be improved with more realistic real world scenarios.

9.7.6 Strengths & Weaknesses

The most important part of the evaluation exercise was to determine the strengths and weaknesses of the research based upon the participating experts' perspectives. Also the drawbacks are equally important as the strengths as they can be used as further research

directions and improvement of the proposed solution and the tool. Dr Robert Hegarty and Dr John Haggerty suggested that the scope of the research was broad and should be narrowed down and focus on specific areas and components in order to achieve exhaustive results if more time were given. Similarly, Dr Kalloniatis and Dr Jones stated that more experiments were needed to strengthen this research further through examining the algorithm on a greater number of real cases. Dr Jones also pointed out one of the limitations of the current tool in terms of its ability to handle timestamp traces only. Such a limitation certainly needs to be addressed and the methodology of SOM implementation ought to be improved so that it can manage different types of input.

9.8 Discussion

The evaluation is an important process that is used to check the validity and contribution of a research project; also in this research, the experts' feedback provides a third party, neutral view from a more experienced perspective and can be useful in establishing the areas in which further investigation is required. The first step of the evaluation was to design a number of open ended questions which were used to obtain experts' feedback and comments on this research. The questions were designed in a way that all the desired aspects of the evaluation were covered, ranging from the worthiness of the research, to its scope, methodology, algorithm, SOM clustering and timeline analysis. In addition, the questions were also framed to determine the strengths and weaknesses of the research.

The selection of evaluators was the next important step which was carefully carried out by making a list of potential participants. These participants were selected mainly from an academic background and consisted of doctors and professors who are experts in the field of digital forensics, as well as possessing years of research and teaching experience in the related areas. The selected participants were then contacted and informed that the evaluation was to check the effectiveness of the study.

Once their consent was received, a convenient time as per their schedule was selected for the interview and discussion. Since most of the participants were spread geographically the interviews were conducted using the combination of the telephone, email and VoIP communications. Although 20 requests were sent out, only 6 agreed to participate in the evaluation process. It is envisaged that more participants could be recruited if more time were

given. Those six participants ranged from senior lecturers to researchers from various universities and institutions both within the UK and abroad.

The feedback given by the experts was recorded and analysed in order to deliver a summary of the different findings and suggestions. In general all experts considered that the research topic was valid and that it provides a scientific contribution to the field of digital forensics. The technique of SOM for data clustering was also appreciated by the experts, although several suggested that the justification and theory for using SOM could be elaborated in more detail, and alternative clustering techniques should also be investigated.

How the research could be improved was also addressed. One of the potential improvements was to use multiple techniques for clustering in order to compare and contrast their performance with the chosen SOM clustering. More datasets should also be used in further research and there should be more than one type of input in order to find out the suitability of the tool in a wider range of scenarios.

The participants were fully agreed in their opinions on the legislative compliance of the suggested tool, since any practical applications of the tool would be somewhat limited without it. Currently, when computer crime cases are brought to court, the trials and conviction are carried out based upon the law of the country in which the crime was committed and is being trialled. If the evidence presented by the AFE does not comply with the legislative requirements, the convictions cannot be carried out and the system would be of little aid to digital forensic examiners.

Linking of the artefacts was considered useful by the majority of the experts as they all visualized the importance of connecting the events back and forth in order to get a complete picture of the scenario. Similarly, attempting to determine whether multiple suspects were involved in a case as well as estimating their technical competency was appreciated by the participants who viewed this as an interesting and useful concept. At the same time, the experts expressed caution that the evidence trail indicator should only be used as appropriate and after consideration because otherwise it could unnecessarily prolong trials. Also, most of the participants were affirmative about the use of the crime index database since having such a record would help in establishing any previous patterns that might exist in crimes or incidents of a similar nature. It would reduce the overall time taken for the analysis of a

particular case; also it can provide help in a situation where an offender repeats an offence as a high level of similarity can be observed through matching within the database.

Regarding the suggestions for improvement and enhancement of the research and the AFE, some of the experts suggested that the focus area of the research should be narrowed as the "researcher is trying too much which introduces a bigger challenge" (Dr Robert Hegarty), so the research needs to be "more specific and concentrate on certain components for a better development" (Dr Haggerty).

The participants all agreed that the research has certainly contributed to scientific knowledge in the field of digital forensics. The framework and the algorithm presented for the development of the automated forensic examiner has a substantial basis and can be used for the development of a tool; this can then be used by the digital forensic experts and "law enforcements in particular are crying out for such a tool because of the heavy overload" (Dr Andy Jones).

9.9 Conclusion

It was imperative to evaluate the AFE architecture and its integrated components from an unbiased perspective in terms of its usefulness, technical capabilities, limitations and the possible enhancements that could improve its usefulness. The chapter describes the entire evaluation process which begins with framing queries seeking to extract the relevant information from the subject matter experts. The questions were designed based upon the potential participants, their academic and professional knowledge and experience in the field of digital forensics. The experts were then contacted and those who expressed their consent were briefed about the research, AFE, the experiments and the results drawn from these experiments. The interviews were conducted at the convenience of the participants.

To summarise, it can be stated that the overall evaluation was quite positive and worth the time and effort by all involved. The level of expertise of all the participants and their critical insight were useful to establish the actual practicality of the research and the AFE approach. It was found that the research can be further continued and improved despite the satisfactory results presented. It is also hoped that this research would be taken further based on the inputs

supplied by the participants, and the shortcomings would be resolved in order to benefit the digital forensic experts and the advancement of digital forensic science.

The future research can focus on the development of certain components (e.g. the Technical Competency and the Visualisation components) of the AFE architecture to enhance their capabilities. In addition, as suggested by some of the participants, the research concept could be further tested with the use of alternative artificial intelligence techniques, especially those within the domain of clustering. As far as legislative aspects are concerned, the main challenge for future research would be the development of a smart tool that is truly universal in its approach since related legislations around the world are not unified, and are even influenced by religion or cultural factors. However, since computer crime is a global phenomenon and is not bound by geographical borders, it is vital that this aspect is considered and addressed in such a way that it preserves rights, regardless of the nationality, culture or religion of the involved parties.

10 Conclusion & Future Work

This chapter concludes the achievements of the research and highlights the limitations which form the basis for further study and research in this area. The research aimed to define and design an advanced approach that is able to automate the analysis process of computer forensic in an efficient and timely manner. This aim was achieved by examining the current state of the art to identify the gap need to be addressed and by thoroughly investigating the most suitable approach to tackle the problem. Also, extensive experiments were devised by using four computer crime cases to prove the defined concept; then, the result was evaluated by experts within the field of digital forensics.

10.1 Achievements of Research

The following are the main achievements of this research:

1. The domain of computer crime was investigated to determine the relationship between the development of technology with the rising rates of computer crimes. The sophistication of computer crime was also researched as well as its impact upon the individuals, corporates and the governmental entities (chapter 2).

2. An appreciable understanding of the current state of the art both procedural and technical. Understanding the context within which forensic tools operate and the importance of ensuing methodical processes are adhered to. This work also provided a basis for understanding the extent to which modern tools and processes operate against current technologies (Chapters 3 and 4).

3. One of the achievements of this research was to clearly identify the challenges that are faced by digital forensic examiners. These challenges have been classified under the three main headings of technical, legal and resource aspects. This study achieved a common ground for the legal and technical experts to share common perspectives. Another achievement of this phase of the research was the completion of a survey that aimed to determine the current as well as future challenges by taking views of experts from diverse backgrounds like academic and forensic backgrounds (Chapter 5).

4. The research was able to successfully demonstrate the use of SOM to cluster data for digital forensic purpose. Both publically available and private cases were examined. Detailed clustering analysis was performed to define the influence of SOM network size as well as the input features on the overall SOM clustering performance (Chapter 6).

5. The AEP was proposed as an algorithm that combined SOM and timeline analysis to provide a robust and refined artefact identification process. Experimental evidence suggests that a very positive outcome was obtained (Chapter 7).

6. The research proposed the Automated Forensic Examiner (AFE) as a novel architecture to automate the forensic examination process. The approach consisted of the AEP which interacts with several components such as technical competency measurement and crime indexing database to convert raw input data into a report that can be read by non-forensic experts. Through the integrated components, the proposed AFE enables the digital investigators to perform the analysis in an automated fashion. From the results attained during the experiments, the AFE is able to fulfil a key requirement for the next generation of the digital forensic tool (Chapter 8).

7. An evaluation for the research as whole and for the proposed AFE was performed. Feedbacks and suggestions from experts in the filed were collected to judge on validity of the researched problem from their perspectives and to evaluate the AFE performance according to its outputs. The strengths and limitations of the AFE were identified and flagged for future research work (Chapter 9).

10.2 Limitations of Research

Despite the achievement of the research, there are areas that need to be taken into account for further improvements. These important limitations are listed below:

1. Considering the research nature, data collection was one of the main barriers due to the sensitivity and confidentiality of the data to be experimented upon.

Obviously, it was not easy gaining access for real cases and subjected them to research purpose. The limited number of cases is a significant barrier to better understanding the effectiveness of the proposed approach.

2. The use of SOM for digital forensic purpose has been demonstrated successfully but the approach needs to be tested using other unsupervised pattern classification techniques to conclude if similar or better results can be achieved.

3. The implemented algorithm solely depends on the timestamps during the data mining. In case of deleted executable files, the AFE may not be an ideal tool to use. The algorithm should be improved to overcome this limitation in order to be more efficient across different cases natures.

4. To have a fair and accurate judgment on the research validity and the approach proposed to tackle the researched problem, only feedback from experts within the field was considered. This fact adds another limitation in terms of large number of participants. In addition, it was relatively difficult to get hold of such experts due to the busy time they always have.

10.3 Future Research

The main achievements and limitations of the approach have been highlighted in the previous sections. Based on these factors a number of opportunities exist for further research and/or enhancement. These are outlined below:

10.3.1 Additional Data Mining & Analysing Algorithm

Although the implemented algorithm was reasonably fast, there is a need to develop the algorithm in a way that it can deal with mining large amounts of data and datasets in an efficient manner. As the data to be analaysed is increasing continuously, the implemented algorithm may not be able to cope-up, unless it is enhanced accordingly – particular when considering the advent of big data.

10.3.2 Improving Technical Competency Measurement

Measuring the Technical Competency of the suspect is an important factor to enable AFE and investigators to understand the range and scope of forensic analyses that need to be

undertaken. Although some experts argued that the technical competency of the suspect would be obvious to the investigator according to the case nature, this may not necessarily be true. Criminals with high technical skills have varieties of interests and their crimes are not limited to hacking or similar nature. Hence, it is vital to define the level of the suspect's technical competency regardless the case nature to avoid missing hidden evidences or spending more time in a case than required.

10.3.3 Future implementation of the AFE

Digital forensic analysis is already a computational intensive task with pre-processing of large images taking many hours to complete. The introductions 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.

10.3.4 Global Legal Compliance

The final aim of any digital forensic examination is to trace the suspect's activities and provide evidence that is sufficient enough to prosecute them. However, this is executed within a legal framework of a certain region through trials and courts. There are different laws and regulations across different countries and states, the mater that leads to justice execution. Therefore, it is imperative that AFE complies with such legislation. Additional functionality is required to control and mandate which process can be undertaken and on which data to ensure compliance.

In conclusion, many digital forensic tools exist for investigating e-crimes. However, these tools are lagging behind and the backlogs or the number of outstanding cases is continuously to increase. This research program has, through a published survey highlighted the need for automating the analysis process, then designed and developed an architecture that is able to automate the analysis process and thus reduce the time taken to analyse a computer crime case.

References

1. AcccessData (2015) "Forensic Toolkit", available: http://accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk, last accessed 30 November 2015
2. Adams, C.W. (2008) Legal Issues Pertaining to the Development of Digital Forensic Tools *Third International Workshop on Systematic Approaches to Digital Forensic Engineering* IEEE, DOI 10.1109/SADFE.2008.17
3. Adderley, R., and Musgrove P.B. (2001) Data mining case study: modeling the behavior of offenders who commit serious sexual assaults, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, August 26-29, San Francisco, CA, USA. ACM, 2001, pp. 215-220, 2001
4. Agarwal. A., Gupta. M., Gupa. S., Gupta. S. C., (2011) Systematic Digital Forensic Investigation Model; *International Journal of Computer Science and Security (IJCSS)*; Volume 5 (1)
5. Aggarwal, S., Duan, Z, Kermes, L., Medeiros, B. (2008) E-Crime Investigative Technologies *Proceedings of the 41st Hawaii International Conference on System Sciences* DOI 1530-1605/08
6. Allbritton, D (2006)"Research Methods", *Depaul University*. Available at: http://condor.depaul.edu/dallbrit/extra/psy241/psy241-lec9-nonexperimental-forweb.ppt (Accessed 15 March 2013)
7. Allen, W. (2005).Computer forensics.*Security& Privacy, IEEE, 3* (4), 59-62.
8. Anon., 2012. On-the-spot digital investigation by means of LDFS: Live Data Forensic System *Mathematical and Computer Modelling* Volume 55 223-240
9. APCO (2012) *ACPO Good Practice Guide for Digital Evidence*. [Online]. Available at: http://www.digital-detective.net/digital-forensics-documents/ACPO_Good_Practice_Guide_for_Digital_Evidence_v5.pdf. Last Accessed 30 November 2015
10. Araste, A. R., Debbabi, M. Sakha, A.AndSaleh, M. (2007) analysing multiple logs for forensic evidence *Digital Investigation* S82-S91
11. 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
12. Ashcroft, J (2001) Electronic Crime Scene Investigation: A guide for first responders [online] Available at https://www.ncjrs.gov/pdffiles1/nij/187736.pdf [20 October 2011]
13. Ayers, D., 2009. A second generation computer forensic analysis system *Digital Investigation* Volume 6, p. S 3 4 – S 4 2
14. Ballabio D., Vasighi M. (2012) A MATLAB toolbox for Self Organizing Maps and supervised neural network learning strategies; *Chemometrics and Intelligent Laboratory Systems*; 118, 24–32
15. Baryamureeba, V. and Tushabe, F. (2006) The Enhanced Digital Investigation Process Model *Asian Journal of Information Technology* 5 (7) 790 – 794
16. Basis Technology (2016) "Autopsy" http://www.autopsy.com/
17. Beebe, N.L. and Clark, J.G. (2007) Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, Digital Investigation 4, no. Supplement 1, 49–54.

18. Beebe, N.C. Stacy, S, D. and Stuckey, D. 2009 Digital forensic implications of ZFS *Digital Investigation* S99-S107 doi:10.1016/j.diin.2009.06.006
19. Beebe, N.L, Clark, J.G., Dietrich, G.B., Ko, M.S. and Ko, D. (2011) Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies, Decision Support Systems, volume 51, Issue 4, pages 732-744
20. Bergold, J. & Thomas, S. (2012) 'Participatory Research Methods: A Methodological Approach in Motion', Qualitative Social Research, vol 13, no 1
21. Betz. D. (2012) Cyberpower in Strategic Affairs: Neither Unthinkable nor Blessed. *Journal of Strategic Studies*. 35.5 689-711.
22. Birk, D. (2011) *Technical Challenges of Forensic Investigations in Cloud Computing Environments*http://www.zurich.ibm.com/~cca/csc2011/submissions/birk.pdf [20 March 2012]
23. Broadhurst, R, Grabosky, P, Alazab, M, Bouhours, B, Chon, S, (2014), 'Organizations and Cybercrime', *International Journal of Cyber Criminology*, 8, (1), 1–20.
24. Bryant, R. P. (2008) *Investigating Digital Crime*. New Jersey: John Wiley and Sons Publications.
25. BSA (2012) "Country Report: United Kingdom". *Business Software Alliance* [Online] Available at: http://portal.bsa.org/cloudscorecard2012/assets/pdfs/country_reports/Country_Report_UK.pdf (Accessed: 12 July 2012)
26. Burks, D. (2011) *Massive Storage Capacities Help to Advance Surveillance Technology. Seagate Technology Paper* [Online] Available at: http://www.seagate.com/files/www-content/product-content/sv35-fam/sv35.5/en-us/docs/massive-storage-advances-surv-tech-tp626-1-1202-us.pdf [30 April 2012]
27. Cabinet Office (2011) The UK Cyber Security Strategy: Protecting and promoting the UK in a digital world. [Online] Available at: http://www.carlisle.army.mil/dime/documents/UK%20Cyber%20Security%20Strategy.pdf (Accessed: 5 March 2012)
28. Cantrell. G., Dampier. D., Dandass. Y., Niu. N., Bogen. C.; (2012) Research towars a Partially-Automated and Crime Specific Digital Triage Process Model; *Computer and Information Science*; 5 (2) 29-38
29. Carnegie Mellon University (2014) *E-crime Watch 2014 - 2014 US State of Cybercrime Survey*; [Online] Available at: http://resources.sei.cmu.edu/asset_files/Presentation/2014_017_001_298322.pdf. Last accessed: 30 November 2015
30. Carpenter, G., (2015) Cyber-Attacks and Terrorism Revealed As Top Emerging Risks For 2015, According To Annual Guy Carpenter Survey. [Online]. Available at: http://www.gccapitalideas.com/2014/11/12/cyber-attacks-and-terrorism-revealed-as-top-emerging-risks-for-2015-according-to-annual-guy-carpenter-survey/. Last accesed: 30 Nov 2015
31. Carrier, B. &Spafford, E.H. (2003) Getting Physical with the Digital Investigation Process.*International Journal of Digital Evidence* 2 (2)
32. Carrier, B. (2002a) Open Source Digital Forensics Tools: The Legal Argument. *Research Report*
33. Carrier, B. (2002b) "Defining Digital Forensic Examination and Analysis Tools", *Digital Forensics Research Workshop*. Syracuse
34. Carrier, B. (2003) Defining digital forensic examination and analysis tools using abstraction layers Vol. 1 (4) [online] Available at http://www.cerias.purdue.edu/homes/carrier/forensics [25 April 2012]

35. Carrier, B., (2003) Defining Digital Forensic Investigation. *International Journal of Digital Evidence*. (1):4
36. Carrier, B (2015) Autopsy, [online] Available at http://www.sleuthkit.org/autopsy [15 September 2015]
37. Carrier, B (2016) "sleuthkit" [online] Available at http://www.sleuthkit.org [04 October 2016]
38. Casey, E. (2004) Tool Reviewed WinHex. *Digital Investigation* 1, 114-128.
39. Casey, E. and Friedberg, S. (2006) Moving forward in a changing landscape *Digital Investigation* 3, 1-2
40. Choenni, R., Leertouwer, E., Busker, T. & Mulder, I. (2011) "The Dark Side of Information Technology: A Survey of IT-related Complaints from Citizens" *Proceedings of the 12th Annual International Digital Government Research Conference*: Digital Government Innovation in Challenging Times, Maryland, USA, June 12-15
41. Chouhan, R. (2014) Cybercrimes: Evolution, Detection and Future Challenges. *IUP Journal of Information Technology*. 10.1:48
42. Cisar, P., Maravic Cisar, S., Bosnjak. S., (2014) Cybercrime and Digital Forensics – Technologies and Approaches, Chapter 42 in *DAAAM International Scientific Book* 2014, pp.525-542, B. Katalinic (Ed.), DAAAM International, Vienna, Austria
43. Clearsy (2016) "B-method, formal method for software development", www.methode-b.com, [06 June 2016]
44. Cohen. F. A., (2009) *Digital Forensic Evidence Examination*, Fred Cohen and Associates, Livermore 3rd Edition
45. Computer Fraud and Security (2012) *Cybercrime Attacks Double In Three Years,*Vol 2012, Issue 10, pp. 1-3
46. Craiger, P. (2010) *Computer Forensics Procedures and Methods*.Handbook of Information Security.John Wiley & Sons.
47. Data Breach Investigations Report (2011) *A study conducted by the Verizon RISK Team with cooperation from the U.S. Secret Service and the Dutch High Tech Crime Unit*www.secretservice.gov/Verizon_Data_Breach_2011.pdf [20 February 2012]
48. DoJ (2001) US Department of Justic. *Electronic Crime Scene Investigation: An On-the-Scene Reference for First Responders.* [Online] Available at: https://www.ncjrs.gov/pdffiles1/nij/227050.pdf Last accessed 16 December 2015
49. Dowland, P.S., Furnell, S.M., Illingworth, H.M. and Reynolds, P.L. (1999) Computer Crime and Abuse: A Survey of Public Attitudes and Awareness.*Computers& Security* 18 (8) 715-726
50. Downing, E. (2011) Cyber Security – A New National Programme. *House of Commons Library.Science and Environment Section*. SN/SC/5832
51. e-fense (2014) products [Online] Available at: http://www.e-fense.com/products.php [2 October 2015]
52. Earnst and Young (2014) *Get Ahead of Cybercrime*; EY's Global Information Security Survey 2014 [Online] Available at: http://www.ey.com/Publication/vwLUAssets/EY-global-information-security-survey-2014/$FILE/EY-global-information-security-survey-2014.pdf. Last Accessed: 30 November 2015
53. EC-Council Press (2012) "Cyber Safety", ISBN-13: 978-1435483712, page 1- 5
54. Eiland, E.E. (2006) Time Line Analysis in Digital Forensics *New Mexico Institute of Mining and Technology*

55. Endicott-Popovsky. B., (2005) *Community Security Awareness Training*; Proceedings of the 2005 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY
56. FBI (2012) *Crime Scene Search* http://www.fbi.gov/hq/lab/handbook/scene1.htm [26 May 2012]
57. Fei, B., Eloff, J., Olivier, M. and Venter, H. (2006) The use of self-organising maps for anomalous behaviour detection in a digital investigation, Forensic Science International 162, no. 1-3, 33–37.
58. Fei, B.K.L., Eloff, J.H.P. Venter, H.S.and Oliver, M.S. (2005) Exploring Data Generated by Computer Forensic Tools with Self-Organising Maps, *Information and Computer Security Architectures Research (ICSA) Group, Department of Computer Science, University of Pretoria, South Africa*
59. Feyereisl, J. and Aickelin, U. (2009) Self-Organising Maps in Computer Security, In Computer Security: Intrusion, Detection and Prevention, Ed. Ronald D. Hopkins, Wesley P. Tokere, pp. 1-30, Nova Science Publishers.
60. Garfinkel, S.L. (2010a) Digital Forensic Research: Next 10 Years *Digital Investigation* S64-S73 doi:10.1016/j.diin.2010.05.009
61. Garfinkel, S.L. (2012b) Digital forensics: XML and the DFXML toolset *Digital Investigation* 1–14 doi:10.1016/j.diin.2011.11.002
62. Garfinkel, S.L. (2013) Digital media triage with bulk data analysis and bulk_extractor. Computers and Security 32: 56-72
63. Gercke. M.; (2010); Challenges in Developing a Legal Response to Terrorist Use of the Internet; *Defence Against Terrorism Review DATR*; 3 (2) 37-59
64. 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).
65. Gordon, S. & Ford, R. (2006) On the definition and classification of cybercrime. *Journal In Computer Virology* 2, 13-20. DOI 10.1007/s11416-006-0015-z
66. Govloop (2012) "Exploring BYOD In The Public Sector", *Research Report by Govloop and Cisco* [Online] Available at: http://api.ning.com/files/E9piBAScXz0wdJb2obngJ6YUzBQI9vU9FyTMb6du9JsCI8BpV2c5u2aZDuLWTGyQ3F9*ce0pKqqZKcURsbISzQ__/BYODfinal_1.pdf (Accessed: 2 April 2013)
67. Grabosky, P. (2004). The global dimension of cybercrime. *Global Crime 6.1* 146-157
68. Gray, J. and Shenoy, P. (2000) Rules of Thumb in Data Engineering.*Microsoft Research: Advanced Technology Division. Proceedings of 16th International Conference on Data Engineering* 3-12
69. Grispos, G., Glisson, W.B. and Storer, T. (2009) *Calm before the Storm: The Emerging Challenges of Cloud Computing in Digital Forensics*. Glasgow University
70. Guidance Software (2015) "Encase Forensic v7", available: https://www2.guidancesoftware.com/products/Pages/encase-forensic/overview.aspx last accessed 30 November 2015
71. Harms, K (2006) Forensic analysis of System Restore Points in Microsoft Windows XP.*Digital Investigation* 3, 151-158
72. Harnish, R. (2015) Cybersecurity in the world of social engineering. *Cybersecurity in Our Digital Lives 2*. 143
73. Hemingway, P & Brereton, N (2001)"What is a Systematic Review", *Hayward Medical Communications*. [Online] Available at: http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/syst-review.pdf (Accessed: 22 March 2013)

74. HHD Software (2015) Free Hex Editor Neo [Online] Available at http://www.hhdsoftware.com/free-hex-editor date [23 July 2015]
75. 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.
76. HTCIA (2010) *Report on Cyber Crime Investigation*. A Report of the International High Tech Crime Investigation Associationhttp://www.htcia.org/pdfs/2010survey_report.pdf [3 May 2012]
77. http://www.pwc.co.uk/services/audit-assurance/insights/2014-information-security-breaches-survey.html. Last accessed 30 Nov 2015
78. Hunton, P (2011a) The stages of cybercrime investigations: Bridging the gap between technology examination and law enforcement investigation. *Computer Law and Security Review* (27) 61-67 doi:10.1016/j.clsr.2010.11.001
79. Hunton, P (2011b) A rigorous approach to formalising the technical investigation stages of cybercrime and criminality within a UK law enforcement environment. Digital Investigation Vol 7 pp 105-113
80. 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.
81. IC3 (2013) *Internet Crime Report 2013*. [Online]. Available at: http://www.ic3.gov/media/annualreport/2012_IC3Report.pdf
82. 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: 15 April 2013)
83. Ieong, R.S.C. (2006) Digital forensics investigation framework that incorporates legal issues *Digital Investigation* 3S 29-36. doi:10.1016/j.diin.2006.06.004
84. IGRE (2012) "Introduction to Cybercrime and Security", *The Institute for Geospatial Research and Education*[Online] Available at: igre.emich.edu/mytc/sites/default/files/ppt/cybercrime.ppt (Accessed: 5 March 2013)
85. Inikpi O. A., Chris O. I. and David S. P. (2011) A New Approach of Digital Forensic Model for Digital Forensic Investigation, (IJACSA) *International Journal of Advanced Computer Science and Applications* 2, 12
86. International Telecommunications Union (2012a) *Percentage Of Individuals Using The Internet*. [Online] Available at: http://www.itu.int/ITU-D/ict/statistics/ [15 April 2012]
87. International Telecommunications Union (2012b) *World Telecommunication and ICT Indicators Database*. [Online] Available at: http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2012/Individuals_Internet_2000-2011.xls
88. ITA (2011) *Royal Decree No 12/2011 Issuing the Cyber Crime Law*. Information Technology Authority: Sultanate of Oman.
89. Johnson, B. & Christensen, L. (2008) *Educational Research: Quantitative, Qualitative, And Mixed Approaches*. Sage Publications: Thousand Oaks, California
90. Kabay, M.E. (2008) A Brief History of Computer Crime: An Introduction for Students *MSIA School of Graduate Studies Norwich University*
91. Karyda, M. &Mitrou, L. (2007) *Internet Forensics: Legal and Technical Issues.* Second International Workshop on Digital Forensics and Incident Analysis
92. Kayacik, H.G, and Zincir-Heywood, A.N. (2006) Using self-organizing maps to build an attack map for forensic analysis, proceeding of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services, article no. 33 doi: 10.1145/1501434.1501474

93. Kazarian, B., & Hanlon, B. (2011). *SMB Cloud Adoption Study Dec 2010 –Global Report What will be the impact of cloud serviceson SMBs in the next 3 years?* White Paper Sponsored by Microsoft; [Online] Available at: http://www.microsoft.com/Presspass/presskits/commsector/docs/SMBStudy_032011.pdf. Last Accessed 30 November 2015
94. Kessler, G.C. (2007) Anti-Forensics and the Digital Investigator.Proceedings of the 5th Australian Digital Forensics Conference, Edith Cowan University, Perth Western Australia,
95. Kierkegaard, S. M. (2005) Cracking Down On Cybercrime Global Response: The Cybercrime Convention. *Communications of the IIMA*Vol 5 Issue 1
96. Kirwan. G,. Power. A.; (2013) *Cybercrime: The Psychology of Online Offendors*; New York, Cambridge University Press
97. Knapp. K J., Boulton. W. R.; 2006; Cyber-Warfare Threatens Corporations: Expansion into Commercial Environments. *Information Systems Management* 23.2:76
98. Kohonen, T (2013) "Essentials of the self-organising maps", ScienceDirect (2013), *Neural Networks* Vol 37, 52–65
99. Kramer. F. D., Starr. S. H., (2009) *Cyberpower and National Security*, Potomac Books Inc
100. Krishnaswamy, A. (2004) "Participatory Research: Strategies and Tools", *Practitioner: Newsletter of the National Network of Forest Practitioners* vol. 22, pp. 17-22
101. Leslie. D. A., (2010) *Legal Principles for Combatting Cyberlaundering*; [Online] Springer Switzerland; Available from: https://books.google.co.uk/books?id=gy8qBAAAQBAJ&pg=PA29&lpg=PA29&dq=curtis+et+al+2010+cybercrime&source=bl&ots=f4_NuQUcX5&sig=rEj8H8IW6wwmctxNsoc9qFtPccw&hl=en&sa=X&ved=0ahUKEwiE5fmBz7rJAhXGWRQKHTB3D6YQ6AEIIzAA#v=onepage&q=curtis%20et%20al%202010%20cybercrime&f=false; Last accessed 30 November 2015
102. 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
103. Mark Reith, Clint Carr and Gregg Gunsch, (2002) An Examination of Digital Forensic Models *International Journal of Digital Evidence* Fall 2002, 1(3)
104. McAfee (2014). *McAfee Internet Security* (Online Download). Available from: http://download.mcafee.com/products/manuals/en-us/MIS_DataSheet_2014.pdf. Last Accessed 30th November 2015
105. McGuire. M., Dowling. S., (2013) Cybercrime: A review of the Evidence Research Report 75 Summary of key findings. [Online] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/246749/horr75-summary.pdf
106. Mee, V., Tryfonas, T. and Sutherland, I., 2006. The Windows Registry as a forensic artefact: Illustrating evidence collection for Internet usage. *Digital Investigation*, Vol 3, 166-173.
107. Mohay, G. (2005) "Technical Challenges and Directions for Digital Forensics", *Proceedings of the First International Workshop on Systematic Approaches to Digital Forensic Engineering* (SADFE'05)
108. Moor's Law (nd) "Moor's Law" [Online] Available at: http://www.mooreslaw.org/
109. Moore, T. (2006) The Economics of Digital Forensics. *Fifth Workshop on the Economics of Information Security*, 26-28 June.

110. Murphey, R. (2007) Automated Windows event log forensics.*Digital Investigation*, Issue 4s, s92-s100.
111. NAO (2013) *Update on the National Cyber Security Programme* [Online]. Available from https://www.nao.org.uk/wp-content/uploads/2015/09/Update-on-the-National-Cyber-Security-Programme-summary.pdf . Last accessed 30 November 2015
112. NCA (2014) Cloud Aware Dynamic Resource Allocation Middleware for Massively Multiplayer Online Games
113. ONS (2014) *Discussion paper on the coverage of crime statistics.* [Online] Available from: file://psf/Home/Downloads/discussionpaperoncoverageofcrimestatistics_tcm77-350307.pdf Last Accessed: 30 November 2015
114. Palmer, G. (2001) *A Road Map To Digital Forensic Research* [online] Available at: http://www.dfrws.org/2001/dfrws-rm-final.pdf [5 May 2012]
115. Palomo, E.J., North, J., Elizondo, D., Luque, R.M. and Watson, T. (2011) Visualization of network forensics traffic data with self-organizing map for qualitative features, Proceedings of the International Joint Conference on Neural Networks, IEEE, San Jose, California, USA, pp. 1740-1747
116. Pant, M. (2015) Participatory Research. Available at: http://www.unesco.org/education/aladin/paldin/pdf/course01/unit_08.pdf (Accessed 5 September 2015)
117. Paraben Corporation (2015) Computer Forensics/ P2C Available at: https://www.paraben.com/p2-commander.html [25 August 2015]
118. Parker, D.B. (2007) The Dark Side. *The Dark Side of Computing: SRI International and the Study of Computer Crime*. IEEE Computer Society
119. Peisert, S., Bishop, M., Karin, S. &Marzullo, K. (2007) Toward Models for Forensic Analysis. Proceedings of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE'07)
120. Perumal (2009) Digital Forensic Model Based On Malaysian Investigation Process *International Journal of Computer Science and Network Security*, 9(8), 38-44
121. Phillippsohn, S. (2001) Trends In Cybercrime - An Overview Of Current Financial Crimes On The Internet *Computers & Security* 20 53-69
122. Pinguelo F.M. and Muller, B.W. (2011) Virtual Crimes, Real Damages: A Primer On Cybercrimes In The United States and Efforts to Combat Cybercriminals *Spring 2011 University Of Virginia* 16(1)
123. Pollitt, (1995) "Computer Forensics: An Approach to Evidence in Cyberspace", in *Proceeding of the National Information Systems Security Conference, Baltimore*, MD, Vol. II, pp.487-491.
124. Pollitt, M. (2007) An Ad Hoc Review of Digital Forensic Models, Vol. 10 (12) in Proceeding of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE'07), Washington. [Online] Available at http://www.ieeexplore.ieee.org/ie15/4155337/4155338/04155349.pdf [5 March 2012]
125. Ponemon (2014). Global Report on the Cost of Cyber Crime. [Online]. Available at: http://www.octree.co.uk/Documents/2014-Global-report-on-the-Cost-of-Cybercrime.pdf. Last accessed: 30 Nov 2015
126. Price Waterhouse Coopers (2011) *Cybercrime: Protecting Against The Growing Threat Global Economic Crime Survey*. [Online] Available at: http://www.pwc.com/en_GX/gx/economic-crime-survey/assets/GECS_GLOBAL _REPORT.pdf (Accessed: 20 May 2012)
127. Price Waterhouse Coopers (2014) *Information Breaches Survey* [online] Available from:

128. Reith, M. Carr. C. Gunsch, G. (2002) an examination of digital forensic model. Department of Electrical and Computer Engineering Air force institute of technology. Wright-Patterson. [online] Available at: http://www.utica.edu/academic/institudes/ecii/ijde/articles.cfm?action [23 April 2012
129. Richet. J-L., 2013, Laundering Money Online: a review of cybercriminals' methods. *arXiv preprint arXiv:1310.2368*
130. Rogers, M., Goldman, J., Mislan, R. and Wedge, T. (2006) Computer Forensics Field Triage Process Model. *Conference on Digital Forensics, Security and Law*
131. Rotich, E. K., Metto, S. K., Siele, L. & Muketha, G. M. (2014). A survey on cyber crime perpetration and prevention: a review and model for cybercrime prevention. *European Journal of Science and Engineering*, 2 (1), 13-28.
132. Roussev, V. and Marziale, L. (2007) Forensic discovery auditing of digital evidence containers.*Digital Investigation*, 4, 88-97
133. Roussev, V., (2011) Building Open and Scalable Digital Forensic Tools.Building Open and Scalable Digital Forensic Tools. 11 *Sixth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering* DOI: 10.1109/SADFE.2011.3
134. Roussev, V., Chen, Y., Bourg, T. & Richard, G.G. (2006) md5bloom: Forensic filesystem hashing revisited. *Digital Investigation*.S82-S90.doi:10.1016/j.diin.2006.06.012
135. Ruan, K., Baggili, I., Carthy, J. &Kechadi, T. (2011) *Survey On Cloud Forensics And Critical Criteria For Cloud Forensic Capability: A Preliminary Analysis*. In 6th annual conference of the ADFSL Conference on Digital Forensics, Security and Law, Richmond, Virginia, USA
136. Ruan, K., Carthy, J., Kechadi, T. &Crosbie, M. (2012) Cloud forensics: An Overview. [Online] Available at: http://cloudforensicsresearch.org/publication/Cloud_Forensics_An_Overview_7th_IFIP.pdf (Accessed: 5 March 2013)
137. Saini. H., Rao. Y. S., Panda. T. C.; (2012) Cyber-Crimes and their Impacts: A Review; *International Journal of Engineering Research and Applications (IJERA)* 2 (2) 202-209
138. Salifu, A. (2008) The Impact of Internet Crime on Development. *Journal of Financial Crime*. 15 (4) 432 – 443.
139. Samuel T. King and Peter M. Chen.BacktrackingIntrusions.*ACM Transactions on Computer Systems*, 23(1):51–76, February 2005.
140. SANS (2012) SANS Institute InfoSec Reading Room, *Computer Forensic Timeline Analysis with Tapestry* [Online] Available at: https://www.sans.org/reading-room/whitepapers/forensics/computer-forensic-timeline-analysis-tapestry-33836 Last access 14 December 2015
141. Sarah Mocas (2004) Building theoretical underpinnings for digital forensics research, *Digital Investigation* 1, 61-68, doi:10.1016/j.diin.2003.12.004
142. Schatz, B., Mohay, G. and Clark, A., 2006. A correlation method for establishing provenance.*Digital Investigation*, Volume 3S, S98 – S107.
143. Scholtz, J. (2010) Towards an Automated Digital Data Forensic Model with specific reference to Investigation Processes.*Proceedings of the 8th Australian Digital Forensics Conference*.
144. Sean, P., Bishop, M., Keith, S.K. Marzullo (2007) Toward Models for Forensic Analysis *Proceedings of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering (*SADFE'07)

145. Selamat, S.R., Yusof, R. and Sahib, S. (2008) Mapping Process of Digital Forensic Investigation Framework, IJCSNS *International Journal of Computer Science and Network Security*, 8 (10)
146. Sridhar, N., Bhaskari, D.L. and Avadhani, P.S. (2011) Plethora of Cyber Forensics *International Journal of Advanced Computer Science and Applications* 2(11) 110-114
147. Stuart, J. and Jon, N. (2003) *Forensic Science: An Introduction to Scientific and Investigative Techniques*. New York: CRC Press
148. Symentec (2015) *Internet Security Threat Report* [Online] Available at: http://www.symantec.com/content/en/us/enterprise/other_resources/21347933_GA_RPT-internet-security-threat-report-volume-20-2015.pdf. Last accessed 30 Nov 2015
149. Tally, G., Sames, D. and Chen, T. (2006) The PhishermanProject :Creating a Comprehnsive data Collection to Combat Phishing Attacks *Journal of Digital Forensic Practice* 1(2)
150. Technical Working Group for Electric Crime Scene Investigation (2001) Electronic Crime Scene Investigation: *A Guide for First Responders*
The 14th IEEE International Symposium on Network Computing and Applications (IEEE NCA14)
151. United Nations (1999) *International review of criminal policy: United nations manual on the prevention and control of computer related crime* www.ifs.univie.ac.at/~pr2gg1 [5 March 2012]
152. UNODC (2012) *UN Comprehensive Study of Cybercrime* [Online] Available at: http://www.unodc.org/documents/southeastasiaandpacific/2012/05/cyber-crime/Bangkok_intro_presentation.pdf. Last accessed 30 November 2015
153. Vacca, J.R. & Rudolph K. (2011) *Systems Forensic Investigation and Response*, Jones and Bartlett Learning, LLC
154. Vlastos, E. & Patel, A. (2008) An open source forensic tool to visualize digital evidence. *Computer Standards & Interfaces*, Issue 30 8-19.
155. Vo¨mel, S. &Freiling, F. C. (2011) A survey of main memory acquisition and analysis techniques for the windows operating system.*Digital Investigation*, Volume 8, 3-22.
156. Wall, D. (2007) *The Transformation of Crime in Cyber Age*. Cambridge: Polity Press
157. Wang, W.B., Huang, M.L., Zhang, J., and Lai, W. (2015) Detecting Criminal Relationships through SOM Visual Analytics, inInformation Visualisation (iV), 2015 19th International Conference on , vol., no., pp.316-321, 22-24 July 2015 doi: 10.1109/iV.2015.62
158. Wilson, N. (2008). Forensics in cyber-space: the legal challenges. In Proceedings of the *1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia Workshop* (pp. 1-6). Adelaide, Australia: ICST
159. Yasin, M., Cheema, A. R. and Kausar, F. (2010) Analysis of Internet Download Manager for collection of digital forensic artefacts.*Digital Investigation*, Issue 7, 90-94.
160. Zhou, J., Heckman, M., Reynolds, B., Carlson, A., and Bishop, M. (2007) Modeling Network Intrusion *ACM Transactions on Information and System Security* 10, 1, Article 4

Appendices

Appendix A: Matlab Code for 4 cases

Appendix B: Publications

- **Challenges to Digital Forensics: A Survey of Researchers & Practitioners Attitudes and Opinions**
  Al Fahdi M, Clarke NL, Furnell SM, Proceedings of ISSA (Information Security South Africa), Johannesburg, 14-16 August, ISBN:978-1-4799-0809-7, 2013
- **Towards an Automated Forensic Examiner (AFE) Based upon Criminal Profiling & Artificial Intelligence**
  Al Fahdi M, Clarke NL, Furnell SM, Proceedings of the 11th Australian Digital Forensics Conference, Perth, Australia, 2-4 December, pp 1-9, ISBN 978-0-7298-0711-1, 2013
- **A suspect-oriented intelligent and automated computer forensic analysis**
  Al Fahdi M, Clarke NL, Li, F, Furnell SM, Digital Investigation 18, 65-76

Appendix C: Questionnaire for the survey study

Appendix D: Evaluation PowerPoint Slides