An Investigation into the use of Data Mining tools to aid forensic investigations

Abstract

As data increases on a yearly basis coupled with information explosion of the World Wide Web since its inception, the amount of data to process has increased drastically. It is imperative to develop techniques to manage this huge data set to extract meaningful information; this process is popularly known as data mining. Data mining is a technique of extracting valuable and useful information, identifying trends, and hidden relationships from large chunks of data that is previously not known. Data mining has found application in various industry, e.g. in healthcare, customer service, business analytics, banking, insurance, forensic and so on. As criminal activities continue to increase around the world, especially cyber-crime, there is a need to develop more efficient forensic tools to analyse this ever increasing data. Data mining has proved to be useful in this regard. This research report aims at investigating how data mining is utilised in the modern world to aid the forensic investigation. The report will also analyse how data mining technique and tools are being utilised globally while also taking into cognisance the type of organisation and individuals utilising these data mining techniques with the sole purpose of identifying trends and providing recommendations for the general forensic community.

Table of Contents

1.0. Introduction

In recent years, advancement in computer technology has been instrumental in solving crimes in various fields of modern life. Nevertheless, technology can also be used as a tool to aid crime. Computer forensics plays a key role in solving various crimes be it cyber-based or non-cyber based. Computer forensics is the application of computer science as well as a technology for the collection and analyses of evidence in a cyber-investigation (Sindu and Meshram, 2012). Computer forensics has played a vital role since the development of personal computers. It has been in existence as far back as 1984 when law enforcement agencies especially the Federal Bureau of Investigation decided to develop programs that will aid in examining computer evidence (Noblett et al., 2000). The explosion of the internet and computer-related crimes has also been a factor for the increase in computer forensic activities. Example of these tools are ProDiscover[1], Forensic Toolkit[2], Encase[3] and so on.

The steady increase in storage media size can cause problems for forensic investigators; analysing a single machine will become cumbersome which in turns will make investigating and analysing of large number of machines more difficult or even impossible (Fei et al., 2006). If we can analyse the computer systems that display suspicious behaviour without analysing all the data on various systems, one can significantly reduce the time and cost of analysing evidence (Fei et al., 2006).

Digital forensics involves the application of scientific techniques to digital media to establish facts from information for judicial review. Digital forensic includes **disk forensics**, **computer forensics**, **network forensics**, **mobile device forensics, device forensics**, **software forensics**, **firewall forensics, database forensics**, **live system forensics** amongst others (Bhat et al., 2010). c. Data mining techniques have vast potentials in forensic science, its models and tools can aid digital forensics professionals, law enforcement agents and professionals to find clues in a more faster and efficient way.

[1] http://www.techpathways.com
[2] http://www.accessdata.com
[3] http://www.guidancesoftware.com

Data preparation/Generation, Data mining and also Data warehousing are three critical features in an investigation (Khobragade & Malik, 2014). Data mining has found application in various fields of study. It can provide insight into human behaviour (Mena, 2003) which as a result can lead to deterring and detecting offenders in a criminal investigation. It has also been helpful in investigative profiling; Computer forensics utilise "after-the-event" or post-mortem to analyse computer crime, this typically involves narrowing down a large search space presented to investigators of such criminal activities. This processing is done on data collected to produce a shortlist of activities that are suspicious; this can then be used by investigators to examine similar evidence in a more detailed fashion (Casey, 2000).

Since most digital evidence is in textual forms such as emails, data mining has been used to analyse different email network for clues that are usually in the form of relationship with data, especially when solving a criminal investigation (Haggerty et al., 2011). Data mining has also found application in fraud detection systems to resolve credit card fraud, insurance fraud, telecommunications fraud and forgery (Thiruvadi and Patel, 2011). This research report will investigate how data mining is utilised in the modern world to aid the forensic investigation.

1.1. Aims and objectives

The project aims at identifying where intelligent data mining tools are used in forensic computing currently, and also provide recommendations on how this use might be extended to other areas and places. The objectives are as follows:

- To collect original data through the process of filling of questionnaires and interviewing specialists in the field of data mining, with a specific focus on those who specialise in data mining in digital forensics
- This project will also investigate how data mining is being utilised in digital forensics, their development, their usefulness and how these may be improved in order to assist a wider implementation especially in places or fields where its application is limited.
- To analyse the data collected from the primary research via questionnaires and interviews to map the current uses of data mining for forensics geographically, by application area and organisationally (which types of organisations are using data mining for digital forensics)

- Identify and analyse a successful example of the use of data mining and from this analyse the criteria which are used in a particular area of digital forensics.
- To provide a set of recommendations into the areas of development from the outcomes of the primary and other secondary research conducted.

1.2. Motivation

The main reason the investigator chose this project is because of the drive to contribute to the relatively new field of digital forensic technology by analysing how it has been deployed in analysing and solving real-world problems, while also identifying useful tools and areas of applications that might prove useful to others.

1.3. The scope of the project

This project will investigate digital forensic tools and its applications amongst forensic experts or researchers majorly in the UK and some other places around the world. The recommendations drawn from the survey will be compared to past literature to identify new tools that are being utilised and are working effectively but have limited adoption; this will be highlighted and suggested for use by the forensic community. The research scope will be limited to only a small fraction of forensic experts/researchers worldwide as the cost and time implication is higher as the sample size increases.

1.4. Report structure

The remaining parts of this report are structured as follows:

Chapter 2 will review forensic tools as well as data mining techniques and tools used currently in forensic investigation. This will also include an analysis of the impact of these data mining techniques and tools.

Chapter 3 is a description of the methodology to be utilised in conducting this research work including the questionnaire and interview questions.

Chapter 4 comprises the analysis and evaluation of the findings from the primary research data to discover trends based on data mining tools utilised as well as the type of organisation amongst others parameters.

Chapter 5 contains valuable recommendations on adopting/improving data mining and forensics tools and processes in the forensic community.

Chapter 6 provides significant conclusions as well as suggested future work for this project.

2.0. Literature Review

In investigating the applicability of data mining tools and technique to aid the forensic investigation, it is essential to conduct a background and historical review of data mining and its application in digital forensics. Emphasis will, however, be focused on two major areas including how data mining can be applied in forensic computing as well as the data mining tools that aid forensic computing. This will be helpful in the analysis of our research which seeks to find out currently utilised data mining tools by comparing our primary research findings to what was observed in the literature.

2.1. Data mining for forensic computing

Data mining technique has boundless potential in the field of forensic science, tools and models and it can be produced to aid or help investigators, law enforcement agents, digital forensic professionals and so on, to find data or clues in a searching exercise in a more productive and faster manner (Khobragade & Malik, 2014). This is a viable option when it comes to analysing a large set of data in an investigative process, e.g. during a crime scene investigation, accounting forensic investigation, cyber investigation and so on, as it offers solution clues that would have taken a very long time to decode without it. Data mining technique using Fuzzy Logic systems, Neural Networks and The Bayesian Belief Network (BBN) has been used in detecting fraud, clustering techniques such as "concept space" has found useful application in crime detection, Neural Network, SVM, Naïve Bayesian and Decision Tree are also useful in email forensic investigation (Appavu & Rajaram, 2007). The rest of this section focuses on how data mining tools, processes and framework are applied in the computer forensic community.

2.1.1. What are computer security and forensics?

Forensic science can be traced back to computer security. Computer security as a field was primarily examined at the beginning of the 1970s. During this period, the technique adopted in the discipline was extensive and more concerned about the improvement of theoretical models rather than an application that have a practical use (Bell & LaPadula, 1973 & Biba, 2000). One of the basis of machine learning applications in computer security is unquestionably spoken about by the work proposed by Denning (1987). From that point forward, a lot of diverse uses of machine

learning has been put forward. Questions about the how security machine learning algorithms are, have been raised over and again (Barenno et al., 2006).

Regardless of these questions, and of the unmistakable comprehension of the development of the security of machine learning calculations (Barenno et al., 2008), both established scientific researchers and tool developers for implementing computer security, now appear to be very much aware of the fact that the role of machine learning systems in the battle against cybercrime. Cases of successful deployment of machine learning exist in various aspects of computer security (Perdisci et al., 2010). More than 30 years of intense effort and billions of dollars have been invested after the first known internet-wide assault occurred in 1988 known as the "Morris Worm" before we arrived at where we are today. Currently, the computer forensic discipline is not younger than computer security because computer security itself has not been in existence for more than 30 years old. By 1984, the FBI Laboratory in collaboration with other law enforcement organisations started creating computer programs to analyse computer proof.

In 1993, the FBI facilitated an International Law Enforcement Conference on Computer Evidence in which 70 representatives of different U.S. government, state, and local law enforcement organisations were in attendance. However, it was not until recently that computer forensic started receiving quality attention from the computer science community. As at the year 2000, the repercussion of cyber-attacks was reported by all the daily papers and TV channels. During the same period, portable PCs, cellular telephones and GPS route navigating system started to overrun the ordinary life of people. Since individuals were turning out to be more acquainted with emails, social networks and instant messaging platforms, it was indeed evident that PCs will have immediately obtained significance regarding the forensic investigation (Ariu et al., 2011).

With the growth of technology, and also the various means of communication, e.g. emails, phones, laptops, there has been an increase also in dataset, hence, forensic investigation typically will require analysis of this ever increasing data when carrying out investigations. It is crucial to develop a technique that would help in analysing this large data set, to find essential clues; data mining renders itself useful in this regards as it helps to find those clues in a relatively short time.

2.1.2. Forensic investigative process/framework

The forensic investigation of any digital evidence is majorly utilised as a post-incident response to an action that can most likely not be characterised as lawful or to an occurrence that does not agree with an organisation's standards and policies. Even though the coverage of physical forensic investigation model has developed during that time, the participation of digital evidence has made its coverage felt recently (Bhat et al., 2010). M Pollitt (1995), recommended a four-stage process which maps acceptance of documented evidence in the court of law to affirm its digital proof, giving a solid base for managing potential digital proof (Pollitt, 1995).

The steps include acquisition, identification, evaluation as well as admission. In 2001, DFRW thought of a framework as seen in figure 1. This is the key framework for all the proposed models that are followed till date.

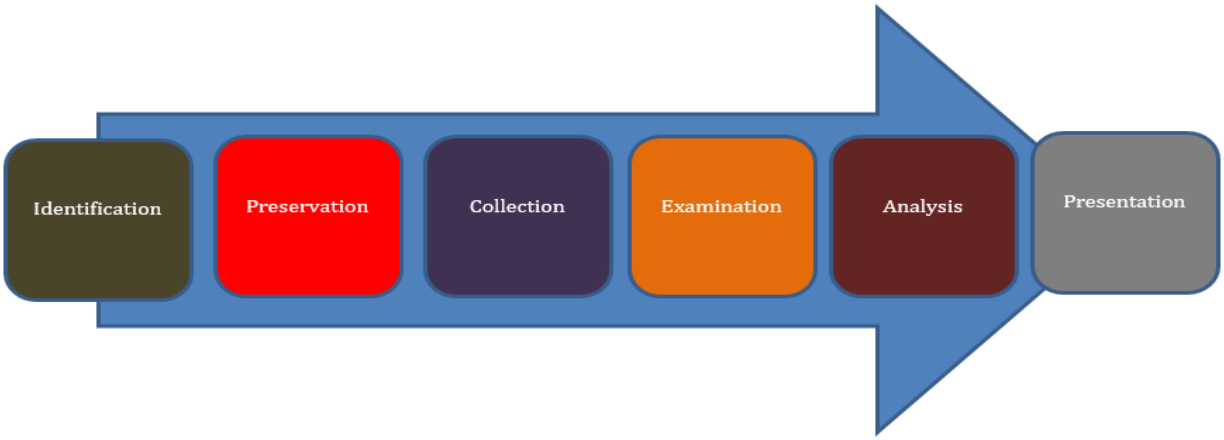

Figure 1: DFRW Forensic Framework

In 2002, Reith et al. suggested a model known as an abstract digital forensic model, taking into account the DFRW model, in which the critical parts of the model included nine stages. The demerits, as cited by the creators, is that the model is excessively broad for practical use, there is no simple or clear technique adopted in testing the model. Also, each subcategory included in the model will make it considerably more cumbersome. In 2006, Kohn, Eloff & Olivier suggested a model, combining the best and vital elements of the considerable number of models that have been proposed to date. The framework is crafted to the point that it can adequately accommodate

any number of the additional stage. The basic model proposed by Freiling & Schwittay (2007); geared toward computer forensic processes and incident response. They permitted an administration situated approach in digital examinations while maintaining the likelihood of a comprehensive forensic investigation (Freiling & Schwittay, 2007). Mohd et al., (2008), recognises the five classes of computer forensic research to include framework, networked environment computer forensics, data detection, reliability, recovery, and lastly, acquisition. The objective of detection and recovery refers to the recognition and recording of digital objects that may have valuable information about an incident. Bhat et al., (2010) emphasised the proposition of an alternative framework for carrying out investigative process on a physical storage device, which expands on the models that have been previously proposed. It likewise proposes and chalks out the execution process for extraction and pre-preparing of data extricated from a flash device/drive.

Even though there are several frameworks that are currently available, it is still difficult to determine the particular forensic framework that is generally adopted in the digital forensic field and the reason for their adoption. This research work also seeks to find out which forensic framework is still most widely accepted via the primary research that will be conducted among forensic professionals in industry and academia. Generally, regardless of the particular framework chosen, data mining would be useful during the analysis of the data stage of any forensic framework adopted.

2.1.3. Forensic tools (including tools using data mining technique)

In analysing and presenting analysed data in a forensic process (via a forensic framework), several forensic tools can be utilised some of which use data mining techniques in conducting the analysis. Table 1 is a list of some data mining tools that can be used for forensic investigation with emphasis on their strengths and weaknesses.

Table 1: Strength and Weaknesses of some Forensic Tools (Lalla & Stephen, 2010)

Tools	General Description	Strength	Weakness
Encase Enterprise Edition v 4.19a (Casey & Stanley, 2004)	Designed for integration with enterprise security architecture, and as such, giving improved access control as well as audit functions while also making it possible for digital investigators to carry out simultaneous system processing on a network	• Can extract more data in comparison to PDIR • Does not change data on a remote system • Data acquisition rate is at 3.5 MB/s • It has a good security system known as SAFE which effectively manages system security concerns. • It can be integrated into an intrusion detection system • It is the most preferred when carrying out enterprise investigations • Provide information about opened files	• Gives a large chunk of possible information usually more than what is necessary • Inability to view data shared on a network, and as such limiting data amount • Needs administrator rights • SAFE system initial reading device makes data acquisition slow
Integrated E-mail Forensic Analysis Framework (Hevner, 2003)	A Java application that can be used to determine email authorship	Can localise information that relates to a suspect, i.e. Email geographical localisation The theoretical foundation is based on text mining, statistical analysis as well as stylometry working with social networking concepts	• There is a prompt for further investigation with Email social networks • Cohesion technique level needs to be improved to get a more reliable result
AutoMiner (Iqbal et al., 2008)	Utilises novel data mining techniques via frequency patterns comparing it to the written print of a person	• Dynamic extraction of write print which is a unique identifier for authorship • Its accuracy rate is between 86-90% • A robust technique for authorship determination	• Accuracy decreases with a corresponding increase in the minimum acceptable threshold interval • Frequency pattern not being conspicuous leads to manual write prints examination
Support Vector Machine and Stylometry (Corney et al., 2002)	Utilised in email authorship determination	• Gives a system approach to determine the raw style markers relative effectiveness • Has foundation on Structural minimisation principle	• Require more experiments in determining authors to style markers sensitivity • No provision of admissible evidence

Tools	General Description	Strength	Weakness
ProDiscover IR 3.5 (Casey & Stanley, 2004)	Crafted to examine one system at a time and it comes in handy when investigating a small number of systems	• Possess optional password and encryption protection • Only displays information that has been verified to be complete • Last accessed date/time stamps are changed while performing some processes	• Needs administrator rights • Possess data acquisition rate of about 5.5MB/s • Inability to view data shared on a network, as such, limiting data amount • Optional password and encryption protection is not enabled by default

While table 2 comprises predominantly commercial applications, there are also alternative open source forensic tools, some of which include (Huebner & Zanero, 2010; Carvey & Altheide, 2011):

- The Sleuth Kit (TSK) with Autopsy is an open source toolkit with a GUI (Graphical User Interface) utilised for the extraction of files.
- AFFLib incorporating Advanced Forensic Format (AFF) (more information about this tool can be seen in John et al., 2010);
- Fiwalk is focused on automation and also quick disk image processing;
- PyFlag, aimed at bringing together the forensic investigation of different data sources conducted via a Python GUI;
- Digital Forensics Framework (DFF), also known as Open Source Digital Investigation Framework from ArxSys; is used predominantly to collect, preserve as well as reveal digital evidence ((Digital-forensic.org, 2015)
- Open Computer Forensics Architecture (OCFA) is a modular design done by the Dutch police for the automation and analysis of large volumes chunks of digital evidence; and
- The WEKA tool is used to analyse data from flash drive. A statistical approach is employed in validating how reliable the pre-processed data are. It also forms a bridge between the digital forensic investigation team and judicial bodies

It is important to note that open source forensic tools have garnered huge usage in the forensic industry especially Sleuth Kit (TSK) with Autopsy. Even though several tools exist in the forensic industry, and previous research work has analysed some of these tools, our research work seeks

to find out which tool has been the most useful and can be adopted by others in the forensic community as well as new tools that are gaining ground in the forensic industry recently.

2.2. Application of data mining in forensic computing

In the previous section, we looked at some forensic and data mining tools. This section will however, will explore the sectors, theoretical concepts and industry where the data mining tools, techniques and algorithm have been extensively used. This sector would be compared to our primary research to identify trends, new industries or systems where data mining has been extensively utilised. This review will cover fraud detection, crime detection and email analysis.

2.2.1 Fraud detection system

Credit card fraud, insurance fraud, check forgery and telecommunications fraud are the predominant types of fraud. Insurance fraud is rampant in automobile and travel. The Uniform Suspected Insurance Fraud Reporting Form which is adapted for use by the NAIC Antifraud Task Force in 2003, replaces the previous Task Force form. This form creates a standard for insurance data fraud for the insurance industry, making it much easier to report and also track fraud. Fraud detection includes three kinds of guilty parties (Baldock, 1997); Criminal offenders, organised criminal offenders that cause major frauds and offenders who carry out fraudulent activities known as soft fraud. This is usually committed by individuals going through financial difficulties.

A fuzzy logic system included the real fraud evaluation policy via the use of optimum threshold values (Altrock et al. 1995). The outcome demonstrated the possibilities of fraud and the reasons why an insurance claim might be false. The system results were a little better than that of the auditors. Another major logic system utilised two ways to copy how fraud experts reason (Cox et al. 1995). Firstly, the discovery model utilises an unsupervised neural network to discover the connections in data and clusters, after which cluster patterns are recognised. Secondly, the fuzzy anomaly detection model uses the Wang-Mendel algorithm, to discover how providers of healthcare carry out fraud against insurance companies. Hot spots methodology conducts a three-stage process (Williams et al. 1997):

- k-means denotes that clustering algorithm for detecting clusters is utilised as a result of other clustering algorithm being more computationally expensive.

- C4.5 algorithm: the subsequent decision tree can be changed over to a rule set and also pruned
- Visualisation equipment for the evaluation of rules, building statistical briefs of the entities linked with each rule.

Williams (1999) expanded the hot spots methodology via the usage of genetic algorithms for rules generation and exploration. Self-Organising Feature Map by Kohonen's was deployed in classifying injury claims in the automobile industry based on the fraud suspicion size (Brockett et al. 1998). The legitimacy of the Feature Map was then assessed utilising a back propagation algorithm as well as feedforward neural networks. From the result, it can be seen that the technique was more dependable and steady when compared to the fraud assessment. Characterisation methods have turned out to be exceptionally successful in detecting fraud (He et al. 1998; Chen et al. 1999) and later on can be combined to classify crime. The distributed data mining model (Chen et al. 1999) uses a reasonable cost model to check C4.5, CART, as well as the naïve Bayesian grouping models. The strategy was put to use in credit card type transactions.

The neural data mining methodology utilises rule-based association rules for mining representative data and also the Radial Basis Function neural network for mining analogue data (Brause et al., 1999). The methodology discusses the importance of using non-numeric data for fraud detection. It was discovered that the outcome of the associated rules increases the accurate prediction. The SAS Enterprise Miner Software (SAS e-insight, 2000) relies on upon association rules, classification technique as well as cluster detection for fraud claim detection. The Artificial Neural Network (ANN) and the Bayesian Belief Network (BBN) study utilised the STAGE algorithm for BBN in backpropagation for ANN and fraud detection (Maes et al. 2002). STAGE greater than one interchanges between two phases of pursuit: running the initial search method against the objective function, and hill climb running for value function optimisation. The outcome demonstrates that BBNs were much faster to train, yet were slower when connected to new cases. The scores are arranged in descending order of the fraud potential. They also generated the detailed rules associated with the fraud claim. FairIsaac prescribed backpropagation neural systems for fraudulent use of credit card (Weatherford et al. 2002). The ASPECT team (Weatherford et al. 2002) concentrated on neural systems to prepare current client profiles and also client profiles histories. A guest's existing profile and the guest profile history are contrasted

to determine whether or not fraud has occurred (Cahill et al. 2002). It is also used to expand on the adaptive fraud detection framework via the usage of fraud score applications in detecting fraud (Fawcett et al. 1997; Ormerod et al. 2003). It used element BBNs also called Mass Detection device to recognise false claims, which then utilised a rule generator known as the Suspicion Building Tool. Fraud types include insurance, internal, credit card and telecommunications fraud. Numerous kinds of fraud detection include insurance, internal, telecoms and credit card fraud detection. Internal fraud detection comprises of fraudulent account reporting by the management (Bell et al. 2000; Lin et al. 2003), and irregular retail dealings/transactions by workers (Kim et al. 2003).

There are four kinds of insurance fraud: crop insurance (Little et al. 2002), home insurance (Bentley, 2000 & Von Altrock, 1997), automobile insurance fraud detection (Phua et al., 2004; Belhadji et al., 2000; Brockett et al., 2002; Viaene et al., 2004 and Stefano et al., 2001), and health insurance (Riedinger & Major, 2002; Yamanishi et al., 2004). A solitary meta-classifier is utilised in choosing the best base classifiers (Phua et al. 2004), and after that, combined with these base classifiers' forecasts to improve cost savings (stacking-bagging). Fraud data set belonging to the automobile insurance industry is utilised to show the stacking-bagging issue. Fraud detection in credit card refers to the screening of credit card applications (Wheeler et al. 2000), or/and well documented credit card transactions (Fan, 2004;Chen et al. 2004; Chiu et al. 2004 ; Foster et al. 2004; Maes et al. 2002; Syeda et al. 2002; Kim et al. 2002;).

Telecommunications subscription data (Moreau et al. 1997; Rosset et al. 1999; Cahill et al. 2002; Cortes et al. 2003), and wireless and wired phone calls (Burge et al. 2001; Kim et al. 2003) are checked. Credit transactional fraud detection was presented by Foster et al. (2004) and also the prediction of bad debts (Ezawa et al. 1996). Past literature majorly focuses on the video-on-demand website (Barse et al. 2003) and also IP-based telecom service provider (McGibney et al. 2003). Online vendors (Bhargava et al. 2003) and online purchasers (Sherman, 2002) can be checked via automated systems. Fraud identification in government parastatals, e.g., tax (Bonchi et al. 1999) and also customs (Shao et al. 2002) has likewise been documented. Generally, these tools have proved to be effective in helping to ameliorate fraud in these different industries.

2.2.2. Crime detection

Another major area forensic science has proved to be useful is in the area of crime detection. Crime according to Webster's' dictionary is an act or a committing of an act that is not permissible, or the negligence or omission of duty which is required by public law and makes violators liable to punishment via that law. A criminal act includes a range of activities such as simple civic duties violation (for example, illegal parking of vehicles) to crimes organised internationally (for example, the 9/11 US attacks).

Crime data mining methodology and techniques

Data mining in the setting of intelligence analysis and crime is a relatively young field. The. Entity extraction has been utilised for automatic identification of individual, vehicle, narcotic drug, address as well as personal properties from the story reports by police (Chau et al., 2002). Clustering technique such as "concept space" has been utilised to relate automatically with diverse objects, for example, persons, associations, vehicles and so on, in crime records (Hauck et al., 2002). Deviation detection has an application in network intrusion detection, fraud detection and other criminal analyses involving the trace of abnormal activities. The classification has discovered an application in email spamming detection and people that send unsolicited emails (de Vel et al., 2001). String comparator has been utilised to get deceptive information.

2.2.3. Email investigation

Emails are one of the most crucial medium used in today's world for communication and are therefore one of the most widely investigated data when looking for clues in solving crimes. The following section will provide an in-depth analysis of email forensic. This will be compared to the result obtained from our primary research in this project.

E-mail analysis tools

Researchers have used already existing state of the art data mining methods, a visualisation technique, machine learning algorithms in the implementation of various frameworks and tools. These tools can be applied in various ways because they have a varied function in investigating cyber-crime. Some of these tools are online based, e.g. UnMask, MET that utilises online e-mail data. Some, however, are offline by nature, e.g. IEFAF, EMT and so on, that utilises offline e-mail

dumps in extracting relevant information useful for the forensic investigation (Sobiya et al., 2012). The Malicious E-mail Tracking (MET) and the benchmarking tools of E-mail Mining Toolkit (EMT) system were developed at Columbia University. These tools utilise data mining techniques to carry out behaviour-based analyses as well as social network analysis (Stolfo et al., 2003 & Stolfo et al., 2006). EMS toolkit provides insight into a user's social network (Hongjun et al., 2008). Visualise Association inside E-mails (VAIE) create data models from emails to categorise them using key word searching techniques (Fanlin et al., 2009). Visualisation techniques have been used for e-mail analysis to give a graphical representation of e-mail data. IEFAF possess features such as computational statistical distribution, email authorship analysis performance and data mining models generation (Rachid et al., 2009).

E-mail author attribution

Authorship analysis is a procedure of looking at the attributes of a written piece to deduce its Authorship. Its roots emanate from a scientific examination known as stylometry; which means the statistical analysis done on literary style. Authorship examination is grouped into three noteworthy fields, Authorship identification, similarity detection as well as authorship characterisation (Khan et al., 2012). Authorship identification decides the Likelihood of a written piece to be produced or written by a particular author through the examination of other writings done by that author. It has found application in a little yet differing number of utilisation territories. Major examples are the identification of authors of literature, forensic analysis in criminal cases and also in program codes (Khan et al., 2012). Authorship analysis has been connected to online messages as of late (Sobiya et al., 2012). Admirable results were acquired regarding email authorship analysis on both collected and across diverse topics as stated in (de Vel et al., 2001 & de Vel, 2000). In another writing, four types of composing style including syntactic, lexical, content specific features and structural; alongside SVM were utilised to recognise the writer of online messages and email (Zheng et al., 2006 & Zheng et al., 2003) which was expanded via genetic algorithm as stated by Jiexun et al in 2006. Stylometric elements consolidated with unsupervised strategies have been utilised for authorship identification and also similarity detection as stated by Abbasi & Chen in 2008. A novel strategy termed as Write-print

utilising frequency pattern mining has been produced as stated by Iqbal et al. in 2008, which was further enhanced utilising clustering technique as stated in Iqbal et al. in 2010.

E-mail classification and clustering

Most email mining undertakings are being proficiently done by utilising email classifications. The assignment of email messages from one category to a pre-determined category set is known as Email classification. The goal of automatic email classification, however, is to carry out this activity on behalf of the user via machine learning strategies. Samples of utilisations are automatic mail arrangement into the spam filter, authors or folders (Sobiya et al., 2012). Four different classifiers (Neural Network, Naïve Bayesian, Decision Tree and SVM) have been utilised to detect suspicious emails (Appavu & Rajaram, 2007). Naïve Bayesian classifier has been utilised for recognising dangers from an organisation's quickly growing email dataset as discussed in Shekar & Imambi in 2008. Different studies have uncovered that SVM has been indicated to be exceptionally robust and effective. Clustering techniques are also used where training data set are not accessible via automatic data categorisation. Clustering technique has been utilised widely for authorship analysis and also text categorisation (Sobiya et al., 2012). A powerful advanced digital text analysis methodology has been discussed by Sergio & et al. in 2010; which depends on clustering based text mining technique.

E-mail social network analysis

Social Network analysis refers to the investigation of communication connections or relationship between individuals. It reveals much information about his/her conduct and circle of individuals (associates, companions, relatives and so on) around him/her that he/she connects with (Sobiya et al., 2012). Social Network has been investigated in (Ryan et al., 2007) by actualising a novel algorithm utilising data mining to distinguish client practices in order to locate correspondences examples between elements in an email accumulation to get social standing. The relationship between individuals has been extracted to find criminal groups as discussed in Rabeah et al. in 2011. Goldberg et al. has likewise investigated social Network investigation in 2008. The investigation utilised recursive data mining with a specific end goal to distinguish as often as possible, frequent occurring groups in online messages (for example, blogs, emails, chats and so

on). Studies have demonstrated that frequent pattern mining methods have been extremely fruitful in this problem domain.

2.3 Other areas where data mining tools have been applied

This report has looked at three major areas of data mining application, however, data mining has found application in other research areas and fields such as accounting forensics (Kovalerchuk et al., 2011), fingerprint mining (Baughman et al., 2010) which also has useful application in crime detection as well as supply chain management which also utilises accounting forensics techniques (Warren & Pearson, 2013).

2.4 Real-world case studies

The previous section has explored areas of forensic application and research work conducted in academia. It is also essential to explore how these tools and techniques have been adopted to solve real-world problems. The following sections are a few examples drawn from industry, majorly from the company PwC and Arthur Andersen LLP, to understand how digital forensic have been utilised in solving problems.

2.4.1 Case study - ENRON CASE

The Enron Case scandal was uncovered in October 2001 which later resulted in the bankruptcy of Enron Corporation; the energy (Li, 2010). The fall of Enron has been portrayed as the biggest failure in the historical backdrop of American capitalism, and its fall had a remarkable effect on the financial market. The organisation's breakdown led investors to lose a lot of cash and workers to lose their jobs in addition to their retirement funds and medical insurance. Furthermore, it resulted in the disintegration of Arthur Andersen LLP, which was the audit organisation that performed both the inside and outside accounting on behalf of Enron Corporation (Bratton, 2015).

The government investigation went on for five years and uncovered the complicated and illegal bookkeeping practices encouraged and carried out by Enron's previous executives. The examinations resulted in 31 terabytes of digital data with the inclusion of data from 130 computers, more than 10 million pages of reports and thousands of emails, winnowing proof that helped deliver convictions of the organisation's top officials, among others (The FBI, 2015).

2.4.2. Case study – State sponsored network intrusion, Act of economic espionage (PwC Cybercrime US Center of Excellence: Case Studies, 2010)

An international energy organisation with its headquarters in the US was reached by the FBI and prompted that their network compromised/endangered through a state sponsor. The systems and network of the company are constantly being invaded by a foreign government. The FBI offered to disclose information to on location people holding a US Government Top Secret leeway. Therefore, the organisation recruited the services of PwC's cleared cybercrime team.

The state sponsor utilised a Persistent and Advanced Network Intrusion technique with the aim of compromising many hundreds of topographically separate system to steal many business deals economic intelligence. The hackers compromised many systems geologically scattered across the Middle East, the United States and South America. PwC worked with the customer to uncover this system and also to use forensic techniques to analyse them.

- **Network Forensics:** PwC gathered log data from a scope of customer system to conduct network forensics investigation. This was done to discover patterns in the communication carried out on the compromised network
- **Malware Forensics:** PwC's owned malware lab was utilised to break down suspected malicious documents and processes discovered on the system. This analysis discovered the kinds of incoming as well as outgoing communication linked with the malware and also functionality

PwC's forensic team brought about the accompanying vital discoveries:

- Discovery of the access methodology
- Finding the method utilised in navigating through the company network environment
- The discovered technique used in stealing the domain admin passwords by the state sponsor
- Confirmation of the attackers' constant remote access to the systems and company network
- The determination that the intruder had at least two years access to the company's environment
- Confirmation of theft of data that has economic value

2.4.3. Case study – Global ATM fraud, Network intrusion, PCI data theft, & PII exposure (PwC Cybercrime US Center of Excellence: Case Studies, 2010)

A Global Fortune 100 organisation encountered a cyber-attack and breach of data which resulted in the loss of Personally Identifiable Information (PII) for a considerable number of clients. Amid the cyber-attack, credit card numbers with their pin numbers were breached. The hackers utilised this data to make imposter ATM cards with the stolen information implanted on the cards. The hackers conveyed the fraudulent ATM cards to people situated in a few urban areas all around the world. In a coordinated fashion, the ATM cards were used to withdraw millions of dollars within a short period.

- **Network and malware forensics:** PwC's primary incident response discovered the primary cause of the network intrusion. Further, PwC uncovered all the systems that were compromised in the environment as well as several harsh, remote access capabilities. This included an examination of found custom malware logs and network traffic. There was also a thorough investigation and review of the organisation's corporate system via penetration testing by the security team.
- **Protected data analysis:** PwC broke down both organised and unstructured data on the compromised computer system for PCI and PII data. After distinguishing data sources having PII, PwC consolidated, de-copied and sorted the unique occurrences of the information. PwC's identified the target the data theft, network intrusion as well as the lacking of security protocols that led to the data breach.
- **Computer Forensics**–PwC safeguarded and also analyse several compromised systems. The forensic analysis located the primary point of intrusion, the systems that have been compromised and also the how/where unrecognised data exfiltration.
- **Malware Forensics**–Twelve custom malware incident was discovered and investigated to find the purpose, capability and functionality. This malware was obscure and undetected even by antivirus technology.
- **Data Discovery & Disclosure**–PwC procured and investigated almost 20 terabytes of data with a specific end goal of recognising delicate client data for notification purpose.

2.5. Gaps in the Forensic Industry

Digital forensics is relatively new (Baryamureeba & Tushabe, 2004 and Reith et al., 2002). Different strategies have been proposed, yet no standard strategy exist in the digital forensic field. A few forensic techniques are utilised with positive results in places where others have not been able to prove events beyond doubts. The reasons for this failure are insufficient resources, the absence of adequate training and funds shortage. Specialists are rare and costly (Allinson, 2004). Also, the absence of a professional association overseeing the activities of specialists has been reprimanded (Allinson, 2004 & Tushabe, 2004); this is still the case today. There is also a lack of consistency in terminologies used within the research community. Furthermore, there are irregularities in the wordings used to explain the different processes (Cohen, 2009 & Cohen, 2010); no critical movement from this issue has been noted as of late.

Recent literature uncovers various proposed systems, models and strategies that have been set up trying to portray an efficient digital examination process formally. In digital forensic investigations, various processes concentrate on the diverse actions performed, for example, data extraction. Others tend to be more concerned with the examination of the data extricated from the digital media. There is no agreed universal forensic framework within the forensic community. From the literature studied, it is a bit obvious that open source tools such as WEKA, and so on, are preferred for academia; this might be because they do not incur any licensing fee, while it is difficult to determine from literature which tools are utilised in other industries. It is also difficult to determine which tools are more common in certain parts of the world. Lastly, the uses of data mining technique as stated earlier is still growing. Recent literature has delved into new fields; one of such notable one is the mobile forensic terrain. Xu et al., (2013) proposed a cloud-based framework called MobSafe to analyse the security of mobile applications using data mining technique. However, the application and testing of this framework has been limited to the Chinese market.

2.6. Summary of Literature

This review has looked at how computer forensic has evolved over the years with an emphasis on the impact of data mining in the field. The researcher has explored forensic frameworks as well as how data mining fits into this framework. The researcher has also explored forensic and data

mining tools such as ProDiscover, Autominer and so on, as well as another open source alternative such as WEKA. Industries and sectors where digital forensics, and data mining tools are utilised e.g. fraud detection, crime detection, email forensics and so on, were also explored. Real-world case studies such as the ENRON case were also explored to highlight how these tools have made an impact in the real-world industry. Other examples from PwC and Arthur Andersen LLC were also studied. Finally, a look at gaps that were observed from the literature study was also done.

3.0. Methodology

3.1. Secondary Research

The **Secondary Research** (literature review) was conducted for this project using materials obtained from the XXX library which includes journals, books and articles, and so on. These materials were read to understand data mining concepts, for the purpose of starting this project. Other digital sources such as IEEE Explore and direct science accessible using the University were also utilised in writing the review. The researcher's supervisor was also very helpful by providing many research materials to study, which formed the core of the literature review.

3.2. Primary research

The forensic investigation research for this project was done using the **mixed methodology,** which included the collection, analysis as well as the integration of both the qualitative and quantitative research. The advantages and disadvantages of utilising the mixed methodology approach are provided in table 2:

Table 2: Advantages and Disadvantages of mixed methodology

Advantages of Mixed Methodology	Disadvantages of Mixed Methodology
• It gives strengths which offset both the quantitative and qualitative weakness for this project. • It provides a more comprehensive and full understanding of the research problem than either qualitative or quantitative approaches alone. • It provides a way of designing better and more context-specific instruments. • It makes it possible to explain how causal processes and findings work.	• The research design could be very complicated. • Consumes more time and resources to plan and implement this type of research. • It might be hard to plan and get one method done by leveraging on the findings of another. • It might be vague to solve discrepancies that occur while interpreting findings.

Quantitative research denotes a systematic investigation using a mathematical, statistics or computational technique (Lisa M, 2008). Quantitative research majorly uses questionnaires to gather raw data, while also utilising specific analysis technique, e.g. mathematical models and so on. Contrariwise, **Qualitative** technique is concerned about the studied utilisation, collection and also empirical material selection which includes personal experiences, case study, interviews, observations and so on, (Willig 2008). It can also be defined as an empirical research process with data other than numbers (Willig 2008). This research project includes both quantitative and qualitative analysis and also identifies the advantages and disadvantages of both. Quantitative and qualitative research method will help maximise the outcome of the survey concerning the forensic investigation. The critical analysis is provided in Table 3:

Table 3: The critical evaluation of quantitative and qualitative research methods

Advantages of Quantitative Research	Disadvantages of Quantitative Research
Quantitative Research is valuable amid the early periods of any research study, when the researcher may have the uncertainty of correctly knowing what will be investigated or the area of concentration. This sort of examination does not require a strict outline to arrange before it begins. This gives the researcher the adaptability to let the study build up more ordinarily. Another good position to the subjective investigation, is that the researcher gains a clear and rich data has intensively formed depictions or visual affirmation, e.g., photographs. This sort of investigation looks at the association and social hugeness and how it impacts individuals,	The investigator of a study using quantitative research is energetically involved through out the research. This gives the researcher a subjective point of view of the study and the associated members. The authority translates the examination according to his or her specific uneven point of view, which skews the collected information. Another demerit is that this examination procedure is to a great degree difficult, and can continue going for an extensive period or even years (Woolf, Tanna, and Karam Singh, 1985)

which is beneficial particularly in sociology (Woolf, Tanna, and Karam Singh, 1985)	
Quantitative research allows the investigator to measure and explore information. The relationship between a free and ward variable is put into consideration in the subtle element. This is precious because the investigator is more focused on the findings of the examination (Miles and Huberman, 1994). Quantitative research can be used to test hypotheses in examinations as a consequence of its ability to evaluate information using insights (Miles and Huberman, 1994).	The essential burden of quantitative examination is that the setting of the study or examination is ignored (Miles and Huberman, 1994). Quantitative research does not inspect things in a trademark setting or discuss the important things to have for different people as it is done by a personal examination (Miles and Huberman, 1994). Another hindrance is that a huge populace example must be inspected; the more the example of people investigated, the more accurately or exact the results will be (Miles and Huberman, 1994).

3.2.1. Selection of participants in the survey

The quantitative and qualitative methods were utilised for different categories of individuals, e.g. individuals working in the forensic departments in large and multi-national companies, people working in forensic departments in Small Medium Enterprises (SMEs) and the students who are doing MSc or PhD in forensic related topics. The main advantages of choosing these categories for the project are provided in below:

- **People working in forensic departments in Large and multi-national companies:** Individuals that are working in multi-national companies or large corporations will provide ideas and techniques utilised in the forensic department in large scale.
- **People working in forensic departments in Small Medium Enterprises (SMEs):** Individuals that are working in SMEs will provide techniques and ideas utilised in the forensic department in small and medium scale enterprise

- **The students who are doing MSc or PhD in forensic related topics:** Students will give an in-depth explanation and information about development and advancement in forensic research.

3.2.2. Approaches adopted for the research

Questionnaires (Online Survey) were utilised in conducting quantitative research, while the interview was adopted in getting the qualitative research done. Table 3 is the critical evaluation of each method with its merits and demerits. The personal interview makes it possible to capture detailed explanation to each forensic question as well as the responder's attitude towards forensic and data mining techniques. Contrariwise, the online survey makes it possible to get general information from diverse individuals that are not geographically located in the same place. This means that it becomes possible to get different viewpoints in different continents and countries regarding forensic and data mining techniques. This project uses the onlinesurveys.ac.uk to design the survey. The positive and negative reasons for using different method is provided below:

Personal Interview:

Positive Reasons:

- It is possible to target a specific group of individuals
- It usually has a higher response rate
- Responders' attitude can be observed

Negative Reasons:

- It consumes time
- It might not be cost effective

Online survey:

Positive Reasons

- It is easy to understand; being the most generally used method of survey collection.
- Easy quantification of received responses.
- It provides the platform to answer in a degree of conformity; which means that respondent can easily pass across their opinion.
- An effective response can be received quickly
- It is a cost effective survey
- It is relatively easy to put together

- Can be put together in a relatively short time in comparison to other data capture approaches
- When done remotely, it can prevent or reduce geographical dependence
- Using survey software makes it possible to analyse survey data using advanced statistical technique to check data reliability, validity and significance and also the capability to analyse multiple variables
- Standardised surveys are usually not prone to diverse kind of errors.

Negative reasons

- Since it is one-dimensional, i.e. it contains around 5 to 7 questions, there is a high possibility of failure in getting the true thoughts and attitude of the respondent.
- Respondents might not be encouraged to provide honest and accurate answers.
- Respondents might not want to provide responses that give a negative picture of them.
- Boring respondents or respondents with memory loss might not be fully aware of the reason why they are giving answers.
- Surveys having closed-ended questions might have a lower rate of validity when compared to other question types.
- Errors in data as a result of non-response to a survey question is possible. Also, the number of people that decides to respond to a particular question may be different from those that decide not to respond. This may result in the creation of bias.
- There is a possibility of the answers to the survey being interpreted differently by the respondent, hence, leading to unclear data, e.g. the option "somewhat agree" may mean different things in various subject context. 'Yes' or 'no' answer options might also be problematic, i.e. people might respond with the answer "no" if the option "only once" is not represented as one of the options.
- Customised surveys are subject to risk associated with having certain error types.

The interview questions and questionnaires utilised by the different categories of individuals are provided in appendix A.

3.2.3. Method of Data Collection

The interviews were to be conducted via Skype, telephone and emails with the merits of using each method. Diverse methods were utilised to receive a maximum response to the forensic and data mining questions. The reasons for using each method are provided below:

- **Skype:** This comes in handy when conducting interviews for people geographically located across the globe, to get sufficient response since it is semi-formal. It also helps save time and cost while also observing the respondents facial gestures and expressions.
- **Telephone:** This medium has the potential to reach many individuals as a result of easy access to phones. This is used for people that are far off and will not prefer the skype method but traditional phone call.
- **Email:** Interviewing via emails is somewhat similar to a survey, but respondents are not restricted to choices in a typical survey format and can hence answer the forensic questions in full without restriction, and they can also do so at their convenience.

The Likert scale forensic and data mining questionnaire was published on the web (link is https://coventry.onlinesurveys.ac.uk/forensic-investigation-survey) and also distributed by hand. The reason for adopting this technique to get responses are stated below:

- **Publishing in Web:** It is relatively easy to get responses from diverse geographical and remote areas across diverse countries.
- **Personal handout distribution in the UK:** This is adequate for the respondent that will be visited face to face in the United Kingdom and given the questionnaire to fill out.

3.2.4. Types of Survey Questions

The two predominant survey questions type includes open-ended and closed-ended questions. Open-ended questions do not have categories or predefined options incorporated. Hence the respondents provide their answers (Lavrakas, 2008). On the other hand, the closed-ended questions limit the respondents' answers to the predefined options given in the questionnaire (Lavrakas, 2008). Both techniques are utilised while designing the questionnaire for the forensic and data mining questions because of the advantages they both offer. The critical analysis of both types is provided in Table 4.

Table 4: Critical Analysis of Closed-ended and open-ended questions (Lavrakas, 2008)

	Advantages	Disadvantages
Closed-Ended Questions	• It is time-efficient • The responses are not difficult to analyse and interpret • Not usually prone to errors in answering questions because options are provided beforehand	• Respondents are limited to options provided. Hence cannot express themselves. • Rapport building is limited in this technique
Open-Ended Questions	• Respondents are not limited. Hence they can express themselves adequately and build rapport. • The answers are usually more specific and detailed	• Time-consuming • Responses are harder to analyse • Prone to errors in responding, e.g. incomplete sentence, grammatical errors, some questions left unanswered, and so on.

3.3. Survey Analysis

An analysis of the responses was conducted for the questionnaire and interview. The online web survey platform was analysed; the survey engine provides a fundamental graphical analysis of the responses with its corresponding percentage range provided. Also, the interviews were also analysed with Microsoft Excel to deduce the corresponding percentage range of responses to each question.

3.3.1. Questions used for Interview and survey

Table 5 contains the questions used in this research project and the reason behind its inclusion.

Table 5: Questions used for the interview and survey

Question Number	Interview Question	Reason for choosing the questions
1	Which continent do you reside?	This question will help to categorise responses based on the continent to be aware of the region of world the responses have emanated.
2	Which country do you work?	This question is useful in knowing the country of origin of the responses.
3	What is the name of your company or school?	This question is useful in categorising response from academia and industry.
4	What sector does your company or school fall under?	This question will categorise the responses received concerning the different industrial or academic sectors.
5	Do you carry out a forensic investigation in your company or your academic area?	The research responses received from the respondent who specified yes will proceed to answer the remaining questions below. Otherwise, they can stop filling the survey.
6	Are you directly involved in any related forensic investigations?	This world will serve as a guide to know the number of respondents that are directly involved in the forensic process, to validate the knowledge of the respondents.

7	Can you explain in more details how you carry out your responsibilities?	This question aims at eliciting detailed and specific information about how the investigative process is conducted.
8	What specific investigative technique or tool do you utilise?	This question will help elicit which tools or technique is more popular amongst the respondents.
9	Who are your targets?	This question will help identify those who utilise forensics and data mining techniques services.
10	How will you rate these investigative tools that you use?	The rating will help identify how the respondents perceive the usability and usefulness of this tools
11	Do these tools have any drawbacks?	The drawbacks of the tools are highlighted via this question.
12	Can you recommend some advance forensic tool that you are aware of that can make your work better but are not currently being used?	This will provide insights into the respondents' knowledge of other better forensic tools that they can take advantage of but are currently not utilising.
13	Are you aware of the current advancement in forensic investigation study and research? If so, can you share some with me?	This question provides more insights into the respondents' knowledge of the current forensic research.
14	What forensic framework do you typically follow (e.g. DFWR)?	This question will provide insights into the respondent's knowledge of which forensic framework is being followed.

15	Do you know the theoretical concepts behind these tools? If yes, can you state them?	This question will help elicit if the researcher understands the in-depth concepts behind the tools or they are only just using the tools without a knowledge of how it works.

3.4. Steps followed for deriving the recommendation

Initially, secondary research was conducted and primary research was later conducted. The primary research results were derived from **statistical** and **graphical analysis**. Also, secondary research will go through **manual analysis** by reading, understanding and summarising followed by **content filtering approach** to elicit key points to compare with the primary research. The **comparison method or approach** was used to evaluate the results received from the secondary and primary research. From this critical analysis, appropriate derivations and recommendations were made.

4.0. Analysis and Evaluation

4.1. Introduction

This section contains an analysis of the questionnaire and interview question responses, for the primary survey questions in the previous chapter. A total of 22 responses were received cutting across both the academic and industrial data mining industry. An analysis and evaluation of the response are provided in the following sections.

4.2. Analysis of Survey Questions

4.2.1. General Demographics

The number of responses of Male and Female to the survey was roughly the same as seen in figure 2. The responses were predominantly from four continents; Asia, North America, Europe and Africa as seen in figure 3. No response was received from South America, Australia or Antarctica because no one was asked to fill the survey from that region. Majority of the response came from Europe, i.e. the United Kingdom with eight responses (36.4%). Majority of the responses were from the academic sector in the United Kingdom. This is closely followed by Africa with seven responses (31.8%) and then Asia and North America with four (18.2%) and three (13.2%) responses respectively. And the responses were majorly from the non-academic industry; two (9.1%) respondent from the oil and gas industry and nine (40.9%) respondent from the consultancy sector. The other sectors included the security sector with two respondents and the agricultural sector with one respondent, making up the remaining 13.6% as seen in figure 5 below. 54.5% of the respondent are full-time workers, while 45.5% work part-time in the forensic industry (see Figure 6).

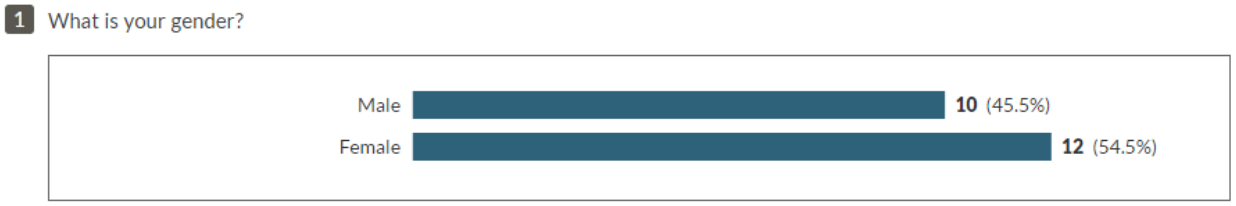

Figure 2: Gender

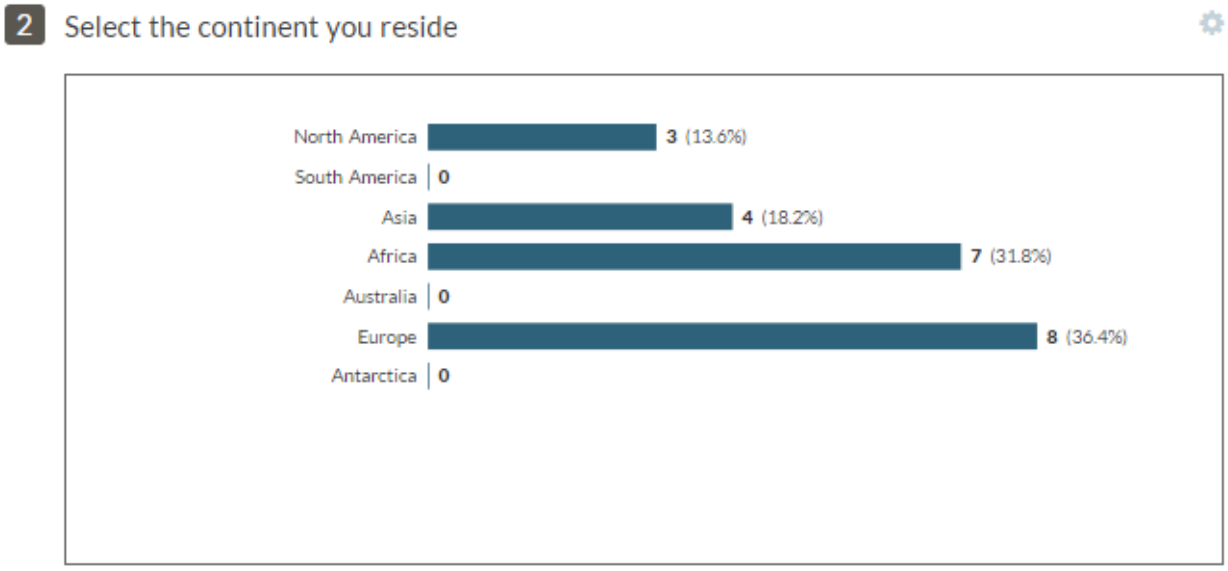

Figure 3: Continent

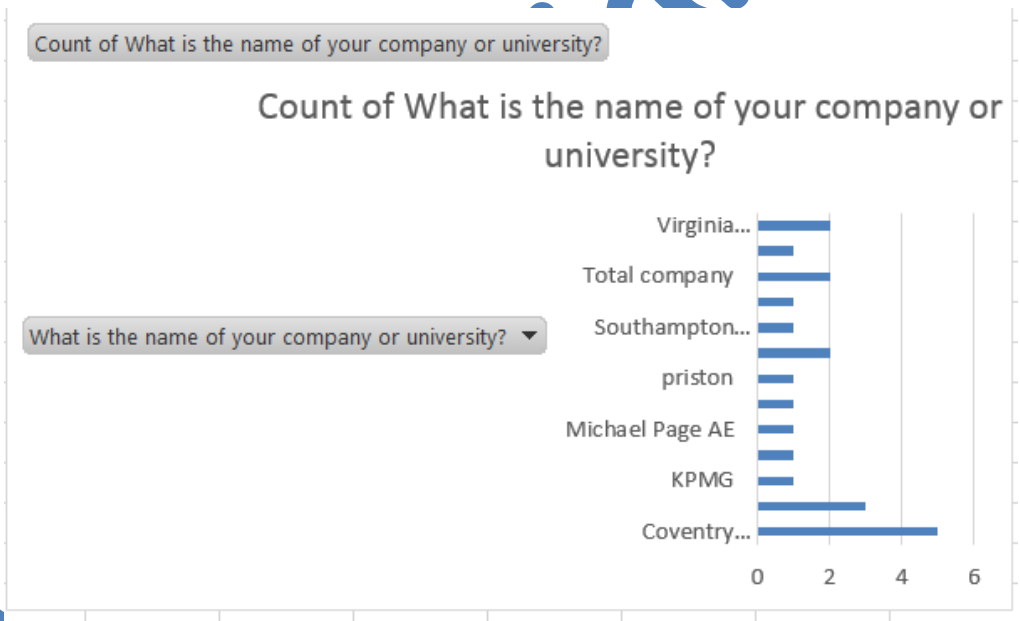

Figure 4: Company or University Name

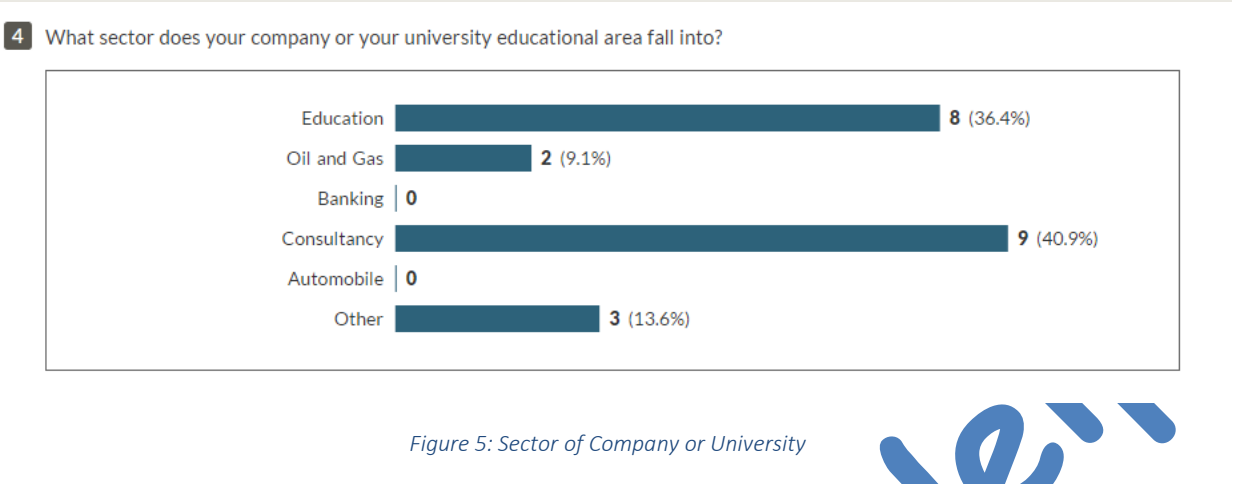

Figure 5: Sector of Company or University

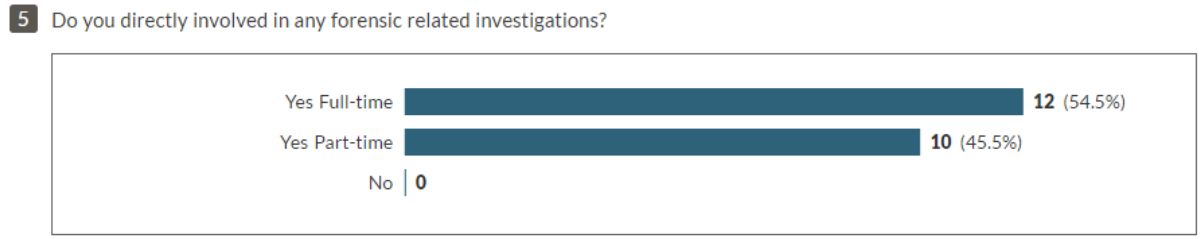

Figure 6: Full Time & Part Time Workers

4.2.2. General Responsibilities

The responsibility carried out by the respondent is as diverse and different as the number of respondents to the questionnaire. However, these responsibilities include:

- Neural network analysis of email linked to fraud detection
- Fixing and prevention of University-related software hacking issues
- Researching into the use of the neural network for text and speech recognition
- Forensic investigation via WEKA, FTK, Imager, Wireshark
- Protection against email and website hacking
- Researching into the evaluation of an intelligent approach to forensic computing
- Supervision of multimedia forensic research
- Researching into the use of Wireshark
- Investigation of client issues and solution provision
- Provision of data analysis and security
- Investigation of a data breach

- Analysis of video evidence
- Formulation, revision and documentation of information security policies, procedures and standards
- Implementing a program to identify and report the use of threat intelligence services and articulate the business benefits it provides
- Implementing a forensics program to identify the required forensics requirements, utilise/leverage current forensics tools and deliver as a pro-active and reactive service
- System auditing and investigation
- Emails analysis and forensic
- Provision of forensic accounting needs
- Data collection, analysis and interpretation.

4.2.3. Which sector (s) uses which forensic/data mining tools?

From the respondents, 30 different tools were provided by the 22 respondent, WEKA tops the chart with 7 respondent, followed closely by ProDiscover IR, Encase, FTK, Autominer amongst others as seen below in figure 7.

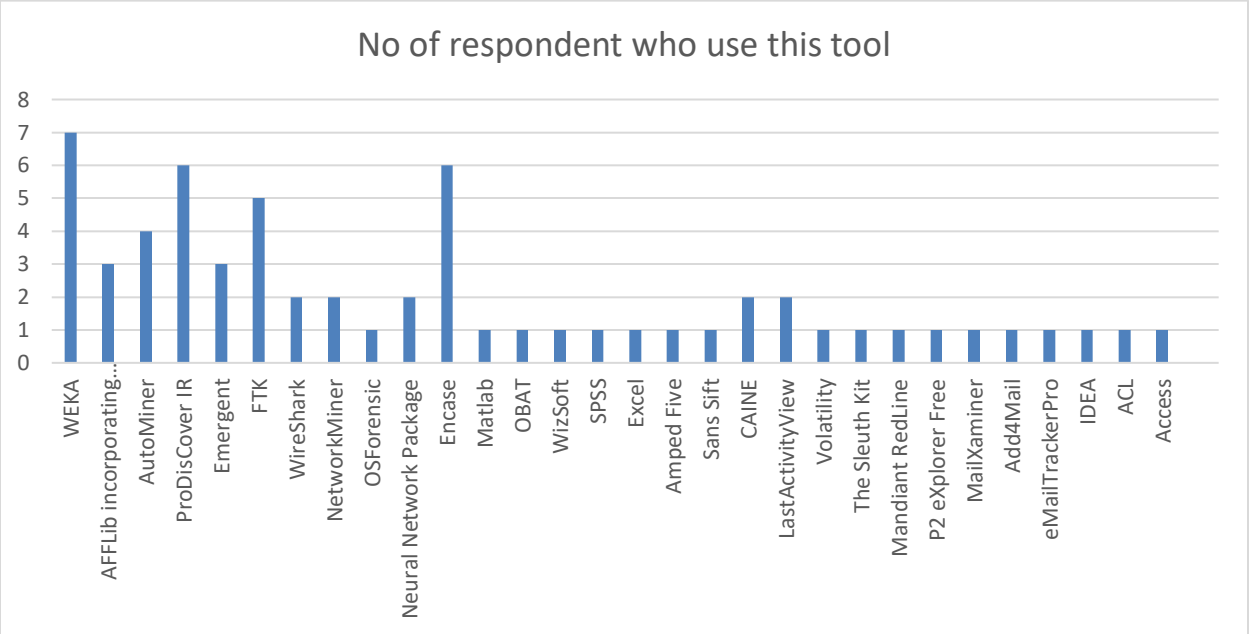

Figure 7: No of respondents who use the tools

However, WEKA has predominant usage in the educational sector. Even though the secondary research in this report does not lay claim to WEKA being predominantly used in the educational sector, it is evident that this is the case, this might be as a result of the fact that the software has been in existence since the late 1990's as well as it has a free GNU General public license. ProDiscover IR, however, has its usage distributed across education, oil and gas and consultancy sector. Encase, and FTK also has usage distributed across the security, education and consultancy sector while Autominer has usage majorly concentrated on the consultancy sector with one respondent, saying it is also utilised in academics. Table 6 has a comprehensive list of these tools as well as the sector they are utilised.

Table 6: Sector(s) each tool are utilized

What is the name of your company or university?	What specific investigative technique or tool do you utilise?	What sector does your company or your university educational area fall into?
Coventry University	WEKA, Emergent	Education
Coventry University	ProDiscover IR, FTK, Wireshark, network miner, OSF	Education
Coventry University	AFFLib incorporating Advanced Forensic Format (AFF)	Education
Coventry University	WEKA, Emergent	Education
Total company	WEKA, Emergent	Oil and Gas
Southampton University	WEKA, Encase, FTK Imager	Education
Ministry of interior ksa	AFFLib incorporating Advanced Forensic Format (AFF), FTK, Encase, Celwbrite	Security
Total company	WEKA, ProDiscover IR, Wireshark, Network miner, Neural Network package	Oil and Gas
priston	AFFLib incorporating Advanced Forensic Format (AFF), WEKA, ProDiscover IR, AutoMiner, Network miner, Neural Network package and Encase	Education
Coventry University	WEKA, Encase	Education
University of Warwick	MatLab	Education

KPMG	OBAT (Online Behavior Analysis Tool) Breach Indicators. IDEA, ACL	Consultancy
KPMG	WizSoft, SPSS, Excel	Consultancy
PwC	ENCASE	Consultancy
PwC	Amped FIVE	Consultancy
Michael Page AE	ProDiscover IR,AutoMiner, Sans Sift, CAINE,LastActivityView, Volatility	Consultancy
Deloitte	AutoMiner, FTK Imager LastActivityView CAINE	Consultancy
Deloitte	ProDiscover IR, The Sleuth Kit (+Autopsy) Mandiant RedLine P2 Explorer Free	Consultancy
Deloitte	ProDiscover IR,AutoMiner, San Sift Incidence Response Encase	Consultancy
Thornton Tomasetti.	MailXaminer Add4Mail eMailTrackerPro	Consultancy
Virginia Department of Agriculture & Consumer Service	IDEA , ACL, Access	Agriculture
Virginia Department of Agriculture & Consumer Service	IDE, Access , ACL	Agriculture

4.2.4. Who are usually the target and what services do they need?

Even though most of the respondent did not give clear target sector of whom they provide forensic services to, from table 7, it is safe to deduce that the major areas where forensics/data mining tools are applied are:

- Accounting Forensic which includes Monetary Fraud Detection
- Email Forensics which also includes Email Fraud Detection
- Crime Forensics
- Computer and Network Forensics

The major application areas in comparison to literature remains the same; no new notable application areas were uncovered from the respondents to the research.

Table 7: Target for forensic work

Who are your target when using forensic tools?
To allow the network to take in email and classify into fraud and non-fraud based on learning frequency of different words
N/A
Generally, they relate to attempts to obfuscate or delete artefacts
Fraudulent activity
fraudulent activity
fraud detection
Find digital Evidence or relative
fraudulent activity
n/a
everything related Wireshark, and how its drawback can be solved
Simulation
Those that need forensic Accounting
Account forensic needs
Any company that needs forensic
Any company that needs forensic video analysis
Companies that require forensic investigation
Companies that need system forensics
Companies that require system auditing
Companies that need a forensic investigation
Email forensic investigation
Agricultural and consumer based companies
Forensic accounting clients

Table 8 below contains a list of the toolset used by the respondent to carry out an investigation as well as the problems they have encountered with these tools, and potential alternatives toolset they are aware of that could aid they work but are not currently utilised. 40.9% of the respondent said they are happy with the toolset they are presently using and cannot see any notable drawbacks while 59.1% stated that drawbacks had been noticed in their current toolset, this can be majorly categorised as seen in figure 8. The predominant problem as highlighted by the respondents are imaging issues, speed and difficulty in use or learning; these are the predominant areas forensic tools provider need focus on for improvement. Also, only 13.6 per cent of the respondents stated that they are not aware of alternative tools that could improve their work while an impressive 76.4% stated that they are aware of alternative tools which they have listed in table 8.

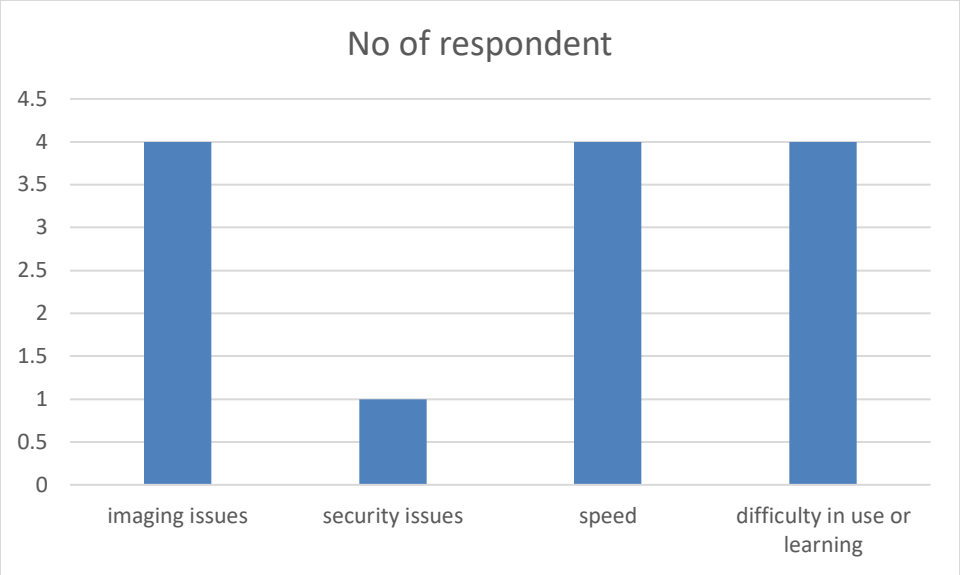

Figure 8: Tools Drawback Categories

Table 8: Toolset drawbacks and alternatives

What specific investigative technique or tool do you utilise?	Do these tools have any drawbacks?	Can you recommend some advance forensic tools that you are aware of that can make your work better but are not currently being used?
WEKA, Emergent	Does not always scale up well	I would recommend that you look at the various intelligent toolboxes offered by Matlab
ProDiscover IR, FTK, Wireshark, network miner, OSF	No	Timeline tools, automated scripts
AFFLib incorporating Advanced Forensic Format (AFF)	In imaging drives, particularly large ones, it takes a great deal of time tying up a workstation	A tool in development we have which uses a Wireshark platform but is enhanced with forensically related modules.
WEKA, Emergent	The tool has problems scaling	None

WEKA, Emergent	the problem with imaging, especially with the large one, it took a long time to tie up the workstation	Matlab has many intelligent tools which are recommended to use
WEKA, Encase, FTK Imager	imaging takes a long time to carry out	interactive classification, decisions tree
AFFLib incorporating Advanced Forensic Format (AFF), FTK , Encase	No	Systenance , WinHex, Bait Tactics
WEKA, ProDiscover IR, Wireshark, Network miner, Neural Network package	No	timeline tools
AFFLib incorporating Advanced Forensic Format (AFF), WEKA, ProDiscover IR, AutoMiner, Network miner, Neural Network package and Encase	Long time to do imaging	neural network package, timeline tools
WEKA, Encase	Security as it is open source	X-way forensics
MatLab	Slow	C#
OBAT (Online Behavior Analysis Tool) Breach Indicators. IDEA, ACL	Most are not so easy to learn	ActiveData
WizSoft, SPSS, Excel	No	None
ENCASE	No	Not at the moment
Amped FIVE	No	dTective

ProDiscover IR,AutoMiner, Sans Sift, CAINE,LastActivityView, Volatility	Some are not user-friendly, e.g. Volatility	Mandiat Redline FTK Imager
AutoMiner, FTK Imager LastActivityView CAINE	No	Oxygen Forensic Suite 2013 Standard Free Hex Editor Neo
ProDiscover IR, The Sleuth Kit (+Autopsy) Mandiant RedLine P2 eXplorer Free	Some tools are not easy to learn	LastActivityView SANS SIFT
ProDiscover IR,AutoMiner, San Sift Incidence Response Encase	No	None
MailXaminer Add4Mail eMailTrackerPro	No	Digital Forensic Framework
IDEA, ACL, Access	Response time is sometimes slow especially with Microsoft Access using Bulk Data	Active Data For Excel
IDE, Access, ACL	Some are difficult to learn	SAS Enterprise Miner Intelligent Data Miner by IBM

4.2.5. How do Forensic users rate their current toolset?

Majority of the respondent have rated the toolset they use highly (see figure 8). This is a clear indication that data mining and forensic tools, in general, have been very useful in its few years of growth.

Visualisation of data

86.4% of the respondents rated data visualisation as “good” on their current toolset. 9.1% said they believe visualisation of data with their current toolset is of an “excellent” standard while only 4.5% of the respondent stated that its current toolset ranks “poor” for visualisation of data.

Speed

86.4% of the respondents also rated their current toolset speed as “good” with only 4.5% respondent rating theirs as “excellent”, while another 9.1% of the respondents were inconclusive whether their toolset speed is good or bad.

Ease of use

63.6% of the respondents rated their current toolset ease of use to be “good”, while 18.2% of the respondents rated their current toolset ease of use as “excellent” with 18.2% of the respondents saying they are inconclusive whether their toolset ease of use is good or bad.

Help System

81.1% of the respondents rated the help system of their current toolset as “good” while 9.1% of the responded rated theirs as “excellent” while another 9.1% of respondent is saying they are not sure if their current toolset is good or bad.

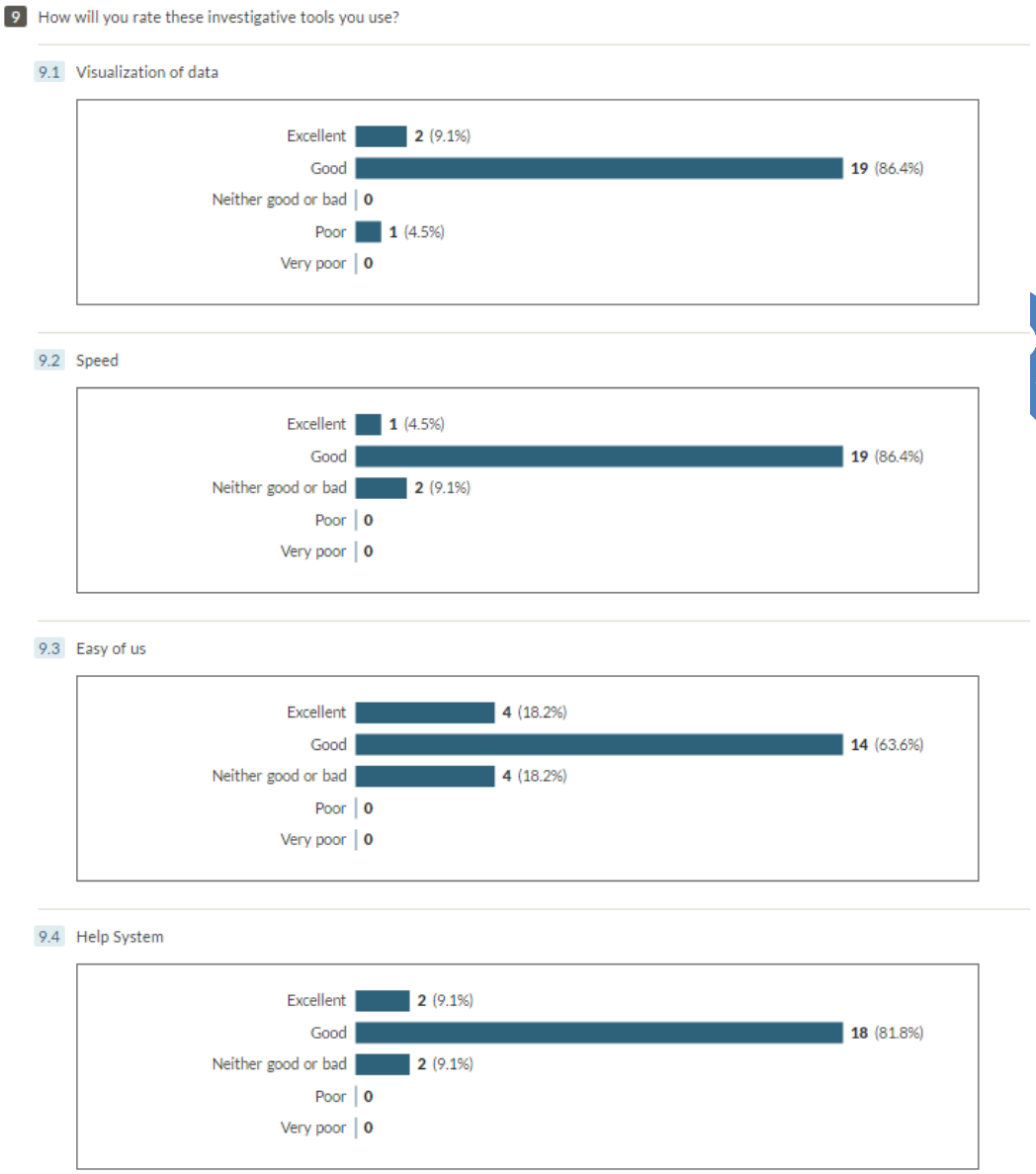

Figure 9: Forensic Tools Rating

4.2.6. Who uses a forensic Framework?

From the respondents, it is evident that there is no clear framework utilised by every respondent in the survey except one respondent that stated that he uses the DFWR framework. This is consistent with claims in the literature that even though several frameworks have been developed, none have cemented its self in the forensic industry.

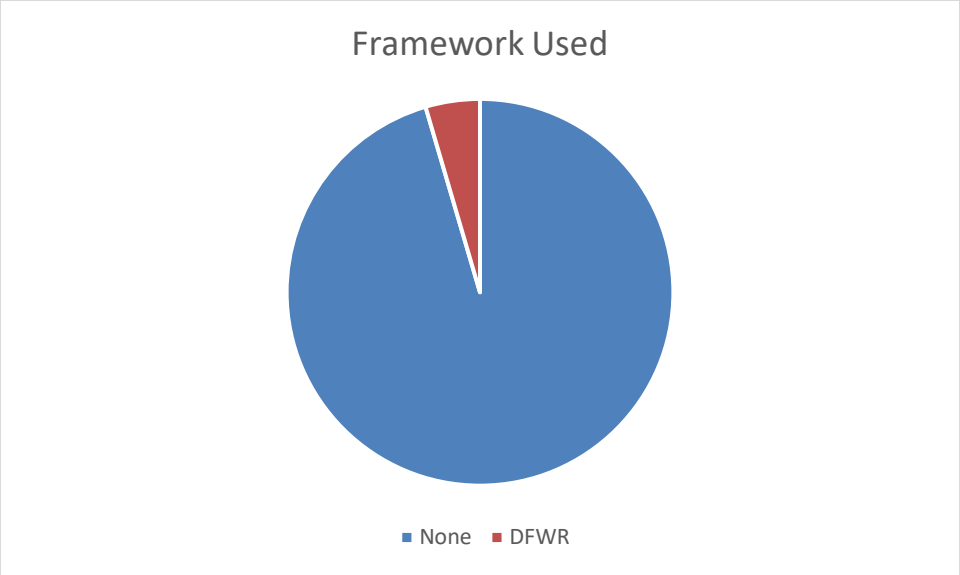

Figure 10: Forensic Framework Used

4.3. Major Findings

From the survey, it is evident that WEKA as a forensic tool is majorly utilised within the academic industry as a research tool. Other dominant tools such as Encase, AutoMiner and FTK are majorly utilised in the industrial sector, and not within the academic sector, industrial sector, e.g. oil and gas, security, agriculture and consulting. Significant areas of forensic application uncovered included accounting forensic (with fraud detection), email forensic (with fraud detection), crime forensics as well as computing and network forensics. The respondents are majorly satisfied with their current forensic toolset when measured under four parameters which included data visualisation, speed, help system and ease of use. However, around 60% of the respondents highlighted some drawbacks from their current forensic toolset which are broadly categorised into "imaging", "speed" as well as problems resulting in "difficulty in learning and using". Also, most respondents are aware of better tools to aid their work, and are not taking advantage of it. Lastly, virtually all of the respondents do not follow any major forensic framework/approach.

5.0. Recommendation

The following recommendations are highly needed in the forensice community to improve the industry at large, after considering the secondary and primary research:

- From both the primary and secondary research, it is obvious that forensic has tremendously benefitted the accounting, emails, crime, computers, network forensics. However, it has majorly been adopted as a mechanism for investigating an incident that has occurred and not before they do. This is one area that the forensic industry should focus on; a kind of defensive forensic mechanism that alerts when something fishy is going on, to limit or ameliorate the damage.
- Even though the current tools utilised have been rated to be adequate, it is clear that some of these tools are difficult to understand and use, while others have scalability and display problems. This is one major area that needs more research focus because data is always increasing and there is always a need to have systems that manage this data management and presentation more efficiently.
- The researcher also observed from the primary research as well as literature, that forensic experts are usually few and they usually have their hands full with work that needs to be done. This could be solved by introducing Forensics as major modules in University computer related, law or accounting related modules or as a whole new discipline at large.
- From primary research and also literature, it is obvious that no framework or methodology is followed within the forensic community, i.e. no standardised approach in solving problems, this might be as a result of lack of professional bodies to manage and propose best practices in the forensic community. This needs to be formulated to monitor and provide relevant training and certification that will increase both the efficiency of forensic workers and also educate more people, and by so doing, increasing the number of forensic professionals in the industry.
- Lastly, it can be seen that most of the respondents, even though they are aware of tools that could potentially increase their productivity, they are presently not taking advantage of these tools. It is recommended that forensic solution providers have a dedicated team of researchers

whose sole duty is to analyse forensic tools, to pick the best tool for the job or encourage forensic team members, especially in the non-academic environment to carry out investigations into the best tool(s) to use.

6.0. Conclusion and Future work

This research work has explored the use of data mining tools capable of aiding the forensic investigation process. This work has uncovered different successful applications of forensic tools and techniques in both the academic and non-academic industry, to aid forensic investigation process. Primary research was carried out to identify new areas of application, but the results did not discover any area of application areas not found in the literature. The major aims and objectives stated in the first chapter of this research work were met with the exception of one of the objectives, which was to map forensic tools based on location. This was not wholly met because of the limitation in the number of respondents to the survey. Additionally, the researcher was unable to get any respondents from Australia and Antarctica. It was not possible to make a significant conclusion from these findings because the data is slightly biased, e.g. most of the academic responses were from the United Kingdom. This research work was conducted to a high standard, and it is worthy to note that the combination of both the secondary research with the result of the primary research lays valid claim and also reflect the current state of the forensic industry in the United Kingdom and around the world.

In conducting this research work, I was able to meet my project milestones because of a personal strict adherence to my timetable and great guidance from my supervisor. However, I had a week delay in between as Iwas unable to conduct the analysis as a result of a limited response to the questionnaires I sent out. Getting the respondents to fill the questionnaire was the most challenging part of the project as I did not take a proactive approach by sending reminders to the people I sent the survey to, so that they can remember to fill it. This is a major challenge in which I have decided to take a different approach in the future. Utilising the survey mechanism for data collection is an acceptable and well-known medium of data collection. Also, the confidentiality of the survey respondents are maintained as no personal information was required from the respondents. The survey was majorly done via a secured survey platform and required access to

the internet, which was readily available to the respondents as this is one of the major resource needed for their work.

Even though this research work has been successful, in the future, the analysis could take a slightly different turn, e.g. in analysing the usability of the tools based on data visualization, speed and so on, it would have been more insightful for the respondents to rate tools individually and not based on the general toolset each respondent uses as was done in this research work. Also, even though the respondents have provided alternative tools that they are currently not utilising, it will also be more insightful to know why they are currently not adopting this tool to improve their work.

7.0. Project Management

These are the set of processes which includes organisation, planning and implementation actions to ensure that this research work was completed on time. This section contains the record of meetings engaged during the project; project timelines, milestones and deadline as well as issues faced and what the researcher did to resolve them.

7.1. Gantt chart

The timelines for the project are provided in figure 11 as seen below. During the project, the researcher followed this timeline judiciously, and the timelines were met. There was no major deviation from the initial plan, except that the project was delayed by one week as a result of not getting enough response from the survey for the analysis. The major milestones in the project included the conclusion of the literature review, methodology, analysis and evaluation, recommendation, conclusion and future as well as writing the final draft of the report. Each milestone was completed before moving to the next one, and they were reviewed weekly to measure how much progress she made.

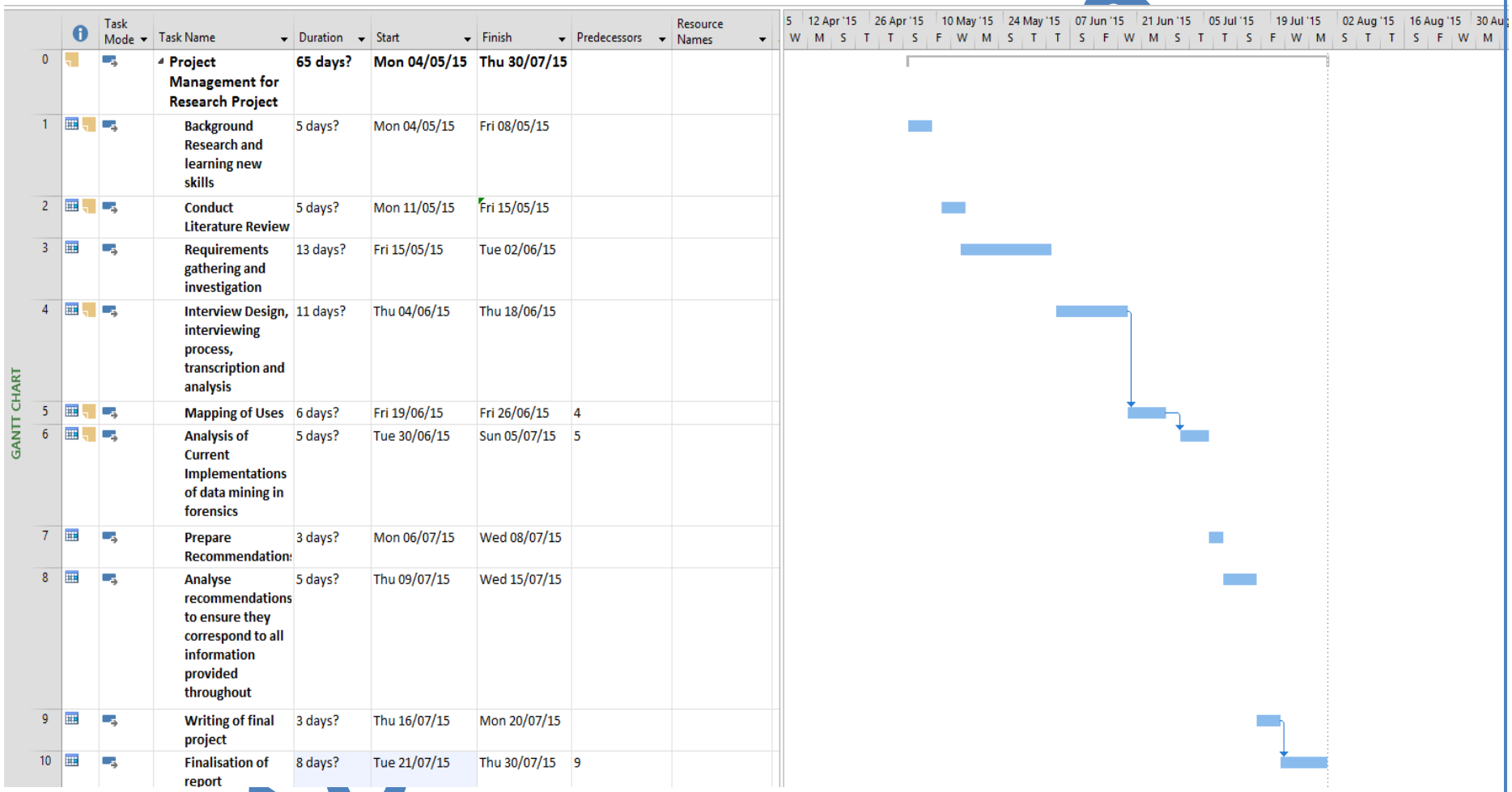

Figure 11: Gantt Chart

7.2. Meetings with supervisor

While carrying out this project work, I had a series of meetings with my supervisor. Figure 12 is a sample meeting sheet, while Table 9 contains a list of meetings I had with my supervisor with discussion points. The "key point at the meeting" as seen in table 9 are the itemised topic or issues that I deliberated and discussed with my supervisor, while the "key action points" represent what needs to be done after the meeting.

Table 9: Supervisor meeting discussion

Meeting date	The key point at the meeting	Key action point
7th May 2015	• Introduction • Explore the structure of the report • Begin Literature	• Start exploring and collecting materials for a literature review • Create a report structure • Read and summarise appropriate papers and books
14th May 2015	• Explore the structure of the report • Continue with literature review • Consider the approach to use for methodology chapter	• Continue collecting literature for literature review • Create a structure of the report • Read and summarise appropriate papers and book chapters as well as adding to report
28th of May 2015	• Continue with the literature survey • Restructure and expand the literature review • Progress made on the comments by the supervisor • Explore the methodology chapter	• Provide the restructured literature review • Devise the structure of methodology chapter • Consider data collection approaches
4th June 2015	• Continue with literature review • Explore methodology chapter	• Provided the restructured literature review • Start to put together the methodology chapter • Review progress so far • Send the current report to supervisor on 14th of June.

		• My supervisor was on vacation between 5th June and 13th June.
16th of July 2015	• Data collection activities • Wait no longer than 24th July to start the analysis • Will explore alternative source to promote questionnaire • Continue with project management chapter	• Consolidate what was done so far • Monitor data collection • Start analysing the response to interviews • Expand the project management chapter • Look at recommendation chapter
30th of July 2015	• Start recommendations chapter • Analysis of data received • Presentation structure. Of course, if you want more or less slides for a particular area that is fine. Remember that you only have 20 mins. ○ Slide 1 – the title of project, name, course ○ Slide 2 – Structure of talk ○ Slide 3 – Aims and objectives ○ Slide 4 and 5 Literature review (main findings) ○ Slide 6-7 Methodology (Diagram of the process and main stages) ○ Slide 8-9 Data analysis (Interesting findings) ○ Slide 10-11 Recommendations and evaluation ○ Slide 12 Did achieve aims and objectives (Table) ○ Slide 13 – critical evaluation and reflection ○ Slide 14 – Future work ○ Slide 15-16 – project management ○ Slide 17 Any Questions • Conclusion chapter structure ○ Did achieve aims and objectives ○ A critical evaluation (Review quality of what you produced)	• Complete the remaining chapter • Focus on the recommendation and conclusion chapter • Supervisor on Holiday 3rd Aug to 9th Aug, so meet next week • My supervisor promised to review the current version on Friday 31st but after that, no email access until 10th Aug • I can look at presentation slides if desired. You can see the second marker if assistance is necessary while I am on vacation

	○ Reflection (How things went, what you did well, what would change write in the first person) ○ Social, ethical, legal and professional issues (Concerned these regarding your project. For example, social factor might be better network performance, so people and institutes can get data quicker. Ethical issue is an intelligent solution that does not say why good suitable and so on) ○ Future work (What will do to extend project)	

Project Study:

Face-To-Face Student Supervision

Student Name	Sanaa Abushofa <abushofs@uni.coventry.ac.uk>
Student ID	5385061
Tutor/ Supervisor/ Director of Studies	Mark Elshaw
Faculty	E and C
Course title and code	
Module code	M08
Date of meeting	7th May 2015
Time of meeting	2pm

Key Points brought to the meeting:

(Include and challenges and actions taken to overcome them)

- Introduction
- Explore structure of report
- Begin literature review

Key Action Points

(Include dates of any deadlines).

- Start exploring and collecting literature for literature survey
- Create structure of report
- Read and summarise appropriate papers and book chapters

- Relate LR Content to project.

Date and Time of next meeting: 2pm 14th May

(Signatures of all those present)

Signature: (Tutor/ Supervisor/ Director of Studies)

Signature: (Student)

Figure 12: Sample meeting sheet

7.3. Project risk analysis

This section considers the risk discovered in relations to this project and possible ways of mitigating them. See table 10.

Table 10: Project risk management

Risk Factor	Reason	Impact	Possible Solution
Health	There might be the possibility of getting sick while conducting the project which might delay the project submission	Medium	A discussion will be held with the supervisor to ask for an extension to submit if necessary
Hardware/software	The computer may experience hardware or software related issue like OS crash	Low	Get a new one or use one from the school library. The project is also backed up regularly on a Google drive account
Funding	To ascertain if project needs funding	Low	The project does not require any additional funding
Safety	To ascertain if the project will need a safety or hazard analysis	Low	The project does not cause any form of damage or safety issues to the environment or individuals
Research related issue	Getting and retrieving materials relevant to the project	Medium	Use university library to source for relevant literature or ask the supervisor for help
Questionnaire	Getting 100 respondents to the survey questions	Medium	Help will be solicited from friends to help distribute the questionnaire, sharing on social media and also asking

			for help from the supervisor.
Report writing issues	Getting the report up to standard readable format	Low	Using grammar and proofreading checkers, engaging the services of a professional proofreader

Contingency Plan: Making sure that the plan is followed will ensure smooth completion of the project before the deadline. Also, allowing an extra one week before the deadline in the Gant Chart will facilitate ease of completion.

7.4. Issues faced during the project

The issues faced during the project are categorised and are contained in the following sections. However, the initially target of respondent was set at 100, as a result of time constraint, the total number of 22 respondents were used in the final analysis.

7.4.1. Issues and solutions faced during the secondary research stage

There are two major issued faced during the stage of my report

- The initial problem encountered while carrying out the secondary research was finding appropriate materials and journals to study and use in the write-up. Even though I was able to get some journals from the IEEE Explore and direct science website from the University, my supervisor was very helpful in providing a rich blend of journals that cuts across both the academic and industry on the application of forensics and data mining technique.
- Another issue I experienced was the synchronisation of the information I read to make meaningful sense, while also justifying the reason for the inclusion of such information; my supervisors' comment was also constructive in this regard, as it helped me structure the flow and justify the reason for each of the information added to the review.

7.4.2. Issues and solution faced during methodology stage

The issues faced during this stage can be broadly categorised into two:

- After publishing the questionnaire and asking a few of the contacts to respond to the survey, the responses were low. After a few days without enough response, I had to send reminder e-mails, text messages and sometimes phone calls to get these individuals to respond to the survey. Also, I observed that because I did not make the fields in the survey form mandatory, some of the fields were not filled; I fixed this issue by making the fields mandatory so that they must be filled before submitting.
- As a result of my initial contact base being small, the number of response I got from the survey was limited. In other to get more response to the survey, I shared the link on social media and via emails with my friends who might know anyone in the forensic community. This was particularly helpful because I observed that I got a few responses to the survey through this medium. I was also more proactive ; I sent reminder emails to those that I sent the survey questions to which was also helpful.

Reference

Abbasi A, Chen H., “Writeprints: a stylometric approach to identity level identification and similarity detection in cyberspace”, ACM Transactions on Information Systems, Vol.26, No.2, Article 7, March 2008.

Ariu, Davide, Giorgio Giacinto, and Fabio Roli. 'Machine Learning In Computer Forensics (And The Lessons Learned From Machine Learning In Computer Security)'. AISec (2011): n. pag. Print.

Baldock, Tony. 'Insurance Fraud'. Australian Institute of Criminology (1997): n. pag. Print.

Barse, E., Kvarnstrom, H. & Jonsson, E. (2003). Synthesizing Test Data for Fraud Detection Systems. Proc. of the 19th Annual Computer Security Applications Conference, 384-395.

Belhadji, E., Dionne, G. & Tarkhani, F. (2000). A Model for the Detection of Insurance Fraud. The Geneva Papers on Risk and Insurance 25(4): 517-538.

Bell, T. & Carcello, J. (2000). A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. Auditing: A Journal of Practice and Theory 10(1): 271-309.

Bentley, P. (2000). Evolutionary, my dear Watson: Investigating Committee based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. Proc. of GECCO2000.

Bezerra, Byron L.D., and Francisco de A.T. de Carvalho. 'A Symbolic Approach For Content-Based Information Filtering'. Information Processing Letters 92.1 (2004): 45-52. Web.

Bhargava, B., Zhong, Y., & Lu, Y. (2003). Fraud Formalization and Detection. Proc. of DaWaK2003, 330-339.

Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D. (1999). A Classificationbased Methodology for Planning Auditing Strategies in Fraud Detection. Proc. of SIGKDD99, 175-184.

Brause, R., Langsdorf, T. & Hepp, M. (1999). Neural Data Mining for Credit Card Fraud Detection. Proc. of 11th IEEE International Conference on Tools with Artificial Intelligence.

Brockett, P., Derrig, R., Golden, L., Levine, A. & Alpert, M. (2002). Fraud Classification using Principal Component Analysis of RIDITs. Journal of Risk and Insurance 69(3): 341-371.

Brockett, Patrick L., Xiaohua Xia, and Richard A. Derrig. 'Using Kohonen's Self-Organizing Feature Map To Uncover Automobile Bodily Injury Claims Fraud'. The Journal of Risk and Insurance 65.2 (1998): 245. Web.

Burge, P. & Shawe-Taylor, J. (2001). An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection. Journal of Parallel and Distributed Computing 61: 915-925.

C.L. Allinson. Legislative and Security Requirements of Audit Material for Evidentiary Purpose. PhD, Queensland University of Technology, http://www.qut.edu.au/, September 2004.

Cahill, M., Chen, F., Lambert, D., Pinheiro, J. & Sun, D. (2002). Detecting Fraud in the Real World. Handbook of Massive Datasets 911-930.

Carvey, H and Altheide, C 2011. Digital Forensics With Open Source Tools. Waltham, MA, Syngress.

Casey, E., & Stanley, A. (2004). Tool review e remote forensic preservation and examination tools. Digital Investigation, 284-297.

Chau, M., Xu, J., & Chen, H. (2002). Extracting meaningful entities from police narrative reports. In: Proceedings of the National Conference for Digital Government Research (dg.o 2002), Los Angeles, California, USA.

Chen, R., Chiu, M., Huang, Y. & Chen, L. (2004). Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines. Proc. of IDEAL2004, 800-806.

Corney, M., Anderson, A., Mohay, G., & De Val, O. (2002). Identifying the Authors of Suspect Email. Computers Security Journal.

Corporation Bedford MA, April 1977. E. Casey (2000). Digital Evidence and Computer Crime. Academic Press.

Cortes, C., Pregibon, D. & Volinsky, C. (2003). Computational Methods for Dynamic Graphs. Journal of Computational and Graphical Statistics 12: 950-970.

Cox, E. (1995), In Goonatilake, S. & Treleaven, P. A Fuzzy System for Detecting Anomalous Behaviors in Healthcare Provider Claims Intelligent Systems for Finance and Business, 111-134. John Wiley.

D. Denning. An intrusion-detection model. Software Engineering, IEEE Transactions on, SE-13(2):222232, February 1987.

D. E. Bell and L. J. LaPadula. Secure computer systems: Mathematical foundations and model. Technical Report M74244 1, MITRE Corporation Bedford MA, May 1973

D.V. Chandra Shekar and S.Sagar Imambi, "Classifying and Identifying of Threats in E-mails – Using Data Mining Techniques", Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol.I, IMECS, 19-21 March 2008, Hong Kong.

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining E-mail Content for Author Identification Forensics. SIGMOD Record, 30(4), 55-64.

Digital-forensic.org, (2015). Free & open Source digital forensics software. [online] Available at: http://www.digital-forensic.org/ [Accessed 7 Jul. 2015].

Ezawa, K. & Norton, S. (1996). Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts. IEEE Expert October: 45-51.

F. C. Freiling, and B. Schwittay, "A Common Process Model for Incident Response and Computer Forensics," Proceedings of Conference on IT Incident Management and IT Forensics, Germany, 2007.

F. Cohen. Digital Forensic Evidence Examination. Fred Cohen & Associates, Second edition, 2009.

F. Cohen. Toward a Science of Digital Forensic Evidence Examination. In Kam-Pui Chow and Sujeet Shenoi, editors, Advances in Digital Forensics VI, volume 337 of IFIP Advances in Information and Communication Technology, pages 17{35. Springer Boston, 2010.

Fan, W. (2004). Systematic Data Selection to Mine Concept- Drifting Data Streams. Proc. of SIGKDD04, 128-137.

Fanlin Meng, Shunxiang Wu, Junbin Yang, Genzhen Yu, "Research of an E-mail Forensic and Analysis System Based on Visualization", Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications, 2009.

Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi., "Mining writeprints from anonymous e-mails for forensic investigation", Digital Investigation, 2010.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Machine Learning, 3.

Fawcett, T., & Flach, P. A. (2005). A response to web and Ting's on the application of ROC analysis to predict classification performance under varying class distributions. Machine Learning, 58(1): 33-38.

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(1-3), pp.33-37.

Food Risk Resource Centre,. 'Mixed Methods Research'. Resourcecentre.foodrisc.org. N.p., 2015. Web. 1 July 2015.

Foster, D. & Stine, R. (2004). Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. Journal of American Statistical Association 99: 303-313.

Given, Lisa M. The Sage Encyclopedia Of Qualitative Research Methods. Los Angeles, Calif.: Sage Publications, 2008. Print.

H Bhat, V., G. Rao, P., R. V, A., Deepa Shenoy, P., K. R., V. and Patnaik, L. (2010). A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application. International Journal of Engineering and Technology, 2(3), pp.313-319.

Haggerty, J., Karran, A., Lamb, D. and Taylor, M. (2011). A Framework for the Forensic Investigation of Unstructured Email Relationship Data. International Journal of Digital Crime and Forensics, 3(3), pp.1-18.

Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H., & Chen, H. (2002). Using Coplink to analyse criminal-justice data. IEEE Computer, 35(3), 30-37.

He H, Wang J, Graco W and Hawkins S.(1997). Application of Neural Networks to Detection of Medical Fraud. Expert Systems with Applications, 13, 329-336.

Hongjun Li, Jiangang Zhang, Haibo Wang, Shaoming Huang, "A Mining Algorithm For E-mail‟s Relationships Based On Neural Networks", International Conference on Computer Science and Software Engineering, 2008.

Huebner, E and Zanero, S (eds) 2010. Open Source Software For Digital Forensics. Springer Heidelberg.

Iqbal F, Hadjidj R, Fung BCM, Debbabi M., "A novel approach of mining write-prints for authorship attribution in e-mail forensics", Digital Investigation 5:pp.42–51, 2008.

Iqbal, F., Hadjidj, R., Fung, B. C., & Debbabi, M. (2008). A Novel Approach of Mining Write-Prints for Authorship Attribution in E-mail Forensics. Digital Investigation , 5 (1), 42-51.

Jiexun Li, Rong Zheng, Hsinchun Chen, “From Fingerprint to Writeprint”, Communications of the ACM, 2006.

John, J.L., Rowlands, I., Williams, P. and Dean, K., 2010. Digital Lives. Personal digital archives for the 21st century. An initial synthesis. A project funded by the Arts and Humanities Research Council, http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf

K. Baughman, A., Van Der Stockt, S. and Greenland, A. (2010). Large Scale Fingerprint Mining.

K. J. Biba. Integrity considerations for secure computer systems. Technical report a423930, MITRE Khobragade, P. and Malik, L. (2014). A Review on Data Generation for Digital Forensic Information using Data Mining. IJCAT International Journal of Computing and Technology, 1(3).

Kim, H., Pang, S., Je, H., Kim, D. & Bang, S. (2003). Constructing Support Vector Machine Ensemble. Pattern Recognition 36: 2757-2767.

Kim, J., Ong, A. & Overill, R. (2003). Design of an Artificial Immune System as a Novel Anomaly Detector for Combating Financial Fraud in Retail Sector. Congress on Evolutionary Computation.

Kovalerchuk, B., Vityaev, E. and Holtfreter, R. (2011). Correlation of complex evidence in forensic accounting using data mining.

Lalla, Himal, and Stephen V. Flowerday. 'Towards A Standardised Digital Forensic Process: E-Mail Forensics'. (2010): n. pag. Print.

Lavrakas, Paul J. Encyclopedia Of Survey Research Methods. Thousand Oaks, Calif.: SAGE Publications, 2008. Print.

Lin, D.-T. Lee, B.-S. P. Lin, S. Shieh, and S. Jajodia, editors, ASIACCS, pages 16–25. ACM, 2006

Lin, J., Hwang, M. & Becker, J. (2003). A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. Managerial Auditing Journal 18(8): 657-665.

Little, B., Johnston, W., Lovell, A., Rejesus, R. & Steed, S. (2002). Collusion in the US Crop Insurance Program: Applied Data Mining. Proc. of SIGKDD02, 594-598.

Lumley, Thomas, and Patrick Heagerty. 'Graphical Exploratory Analysis Of Survival Data'. Journal of Computational and Graphical Statistics 9.4 (2000): 738. Web.

M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar. Can machine learning be secure? In F.-C.

M. Barreno, P. L. Bartlett, F. J. Chi, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, U. Saini, and J. D. Tygar. Open problems in the security of learning. In D. Balfanz and J. Staddon, editors, AISec, pages 19–26. ACM, 2008.

M. Goldberg, M. Hayvanovych, A. Hoonlor, S. Kelley, M. Ismail, K. Mertsalov, B. Szymanski and W. Wallace, “Discovery, Analysis and Monitoring of Hidden Social Networks and Their Evolution”, Technologies for Homeland Security, IEEE Conference, pp.1-6, 2008.

M. Kohn, J. Eloff, and M. Oliver, “Framework for a Digital Forensic Investigation,” Proceedings of Information Security South Africa from Insight to Foresight Conference, South Afrika, 2006.

M. Pollitt, “Computer Forensics: An Approach to Evidence in Cyberspace”, Proceedings of the National Information Systems Security Conference, Baltimore, pp. 487-491, 1995.

M. Reith, C. Carr and G. Gunsch, “An Examination of Digital Forensic Models,” International Journal Digital Evidence, vol. 1, no. 3, 2002.

M. Reith, C. Carr, and G. Gunsch. An Examination of Digital Forensic Models. International Journal of Digital Evidence, 1(3):1{12, 2002.

Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B. (2002). Credit Card Fraud Detection using Bayesian and Neural Networks. Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies.

Magnify (2002). FraudFocus Advanced Fraud Detection, White Paper, Chicago.

Magnify (2002). The Evolution of insurance Fraud Detection: Lessons learnt from other industries, White Paper, Chicago.

Major, J. & Riedinger, D. (2002). EFD: A Hybrid Knowledge/Statistical-based system for the Detection of Fraud. Journal of Risk and Insurance 69(3): 309-324.

McGibney, J. & Hearne, S. (2003). An Approach to Rules-based Fraud Management in Emerging Converged Networks. Proc. Of IEI/IEEE ITSRS 2003.

Mena, J. (2003). Investigative data mining for security and criminal detection. Butterworth Heinemann.

Miles, Matthew B, and A. M Huberman. Qualitative Data Analysis. Thousand Oaks: Sage Publications, 1994. Print.

Mohd Taufik Abdullah, Ramlan Mahmod, Abdul A. A. Ghani, Mohd A Zain and Abu Bakar Md S, “Advances in Computer Forensics,” International Journal Of Computer

Moreau, Y. & Vandewalle, J. (1997). Detection of Mobile Phone Fraud Using Supervised Neural Networks: A First Prototype. Proc. of 1997 International Conference on Artificial Neural Networks, 1065-1070.

Noblett, M., Pollitt, M. & Presley, L. (2000). Recovering and examining computer forensic evidence. Forensic Science Communications, vol. 2, no. 4.

Olivier de Vel, “Mining E-mail Authorship”, KDD-2000 Workshop on Text Mining, August 20, Boston, 2000.

Ormerod T., Morley N., Ball L., Langley C., and Spenser C. (2003). ‘Using Ethnography To Design a Mass Detection Tool (MDT) For The Early Discovery of Insurance Fraud’, Computer Human Interaction, April 5-10, Ft. Lauderdale, Florida.

Phua, C., Alahakoon, D. & Lee, V. (2004). Minority Report in Fraud Detection: Classification of Skewed Data, SIGKDD Explorations 6(1): 50-59.

PwC CybercrimeUS Center of Excellence: Case Studies. (2010).

R. Khan, S., M. Nirkhi, S. and V. Dharaskar, R. (2012). Mining E-mail Content for Cyber Forensic Investigation. UACEE International Journal of Computer Science and its Applications, 2(2).

R. Perdisci, W. Lee, and N. Feamster. Behavioral Clustering of HTTP-Based Malware and Signature Generation Using Malicious Network Traces. In NSDI, pages 391–404. USENIX Association, 2010.

Rabeah Al-Zaidy, Benjamin C. M. Fung, Amr M. Youssef, “Towards discovering criminal communities from textual data”, Proceedings of the 2011 ACM Symposium on Applied Computing, 2011.

Rachid Hadjidj, Mourad Debbabi, Hakim Lounis, Farkhund Iqbal, Adam Szporer, Djamel Benredjem, “Towards an integrated e-mail forensic analysis framework”, Digital Investigation 5, pp.124–137, 2009.

Recommender-systems.org,. 'Hybrid Recommender Systems | Recommender Systems'. N.p., 2015. Web. 27 June 2015.

Rosset, S., Murad, U., Neumann, E., Idan, Y. & Pinkas, G. (1999). Discovery of Fraud Rules for Telecommunications - Challenges and Solutions. Proc. of SIGKDD99, 409-413.

Ryan Rowe, German Creamer, Shlomo Hershkop and Salvatore J Stolfo, “Automated Social Hierarchy Detection through E-mail Network Analysis”,Joint 9th WEBKDD and 1st SNAKDD Workshop ‘07 August 12, 2007, San Jose, California, USA.

S S.Appavu alias Balamurugan, Dr.R.Rajaram, “Data mining techniques for suspicious e-mail detection: A comparative study”, IADIS European Conference Data Mining, 2007.

SAS e-Intelligence (2000). Data Mining in the Insurance industry: Solving Business problems using SAS Enterprise Miner Software, White Paper.

Sas.com,. 'Statistical Analysis - What Is It?'. N.p., 2015. Web. 27 June 2015.

Sergio Decherchi, Simone Tacconi, Judith Redi, Fabio Sangiacomo, Alessio Leoncini and Rodolfo Zunino, “Text Clustering for Digital Forensics Analysis”, Journal of Information Assurance and Security 5 (2010),pp.384-391.

Shao, H., Zhao, H. & Chang, G. (2002). Applying Data Mining to Detect Fraud Behavior in Customs Declaration. Proc. of 1st International Conference on Machine Learning and Cybernetics, 1241-1244.`

Sherman, E. (2002). Fighting Web Fraud. Newsweek, June 10.

Sindu, K. and Meshram, B. (2012). A Digital Forensic Tool for Cyber Crime Data Mining. IRACST – Engineering Science and Technology: An International Journal (ESTIJ), Vol.2 (No.1).

Sobiya R. Khan, Smita M. Nirkhi, R. V. Dharaskar, “E-mail Mining for Cyber Crime Investigation”, Proceedings of International Conference on Advances in Computer and Communication Technology, pp.138-141, February 2012.

Stefano, B. & Gisella, F. (2001). Insurance Fraud Evaluation: A Fuzzy Expert System. Proc. of IEEE International Fuzzy Systems Conference, 1491-1494.

Stolfo S.J., Hershkop S., Ke Wang, Nimeskern O., “EMT/MET: systems for modeling and detecting errant e-mail”, Proceedings of DARPA Information Survivability Conference and Exposition, 2003.

Stolfo S.J., Hershkop S., Ke Wang, Nimeskern O., Chia-Wei Hu, “Behavior-Based Modeling and Its Application to E-mail Analysis”,ACM Transactions on Internet Technology, Vol. 6,No. 2, May, Pages 187–221, 2006.

Syeda, M., Zhang, Y. & Pan, Y. (2002). Parallel Granular Neural Networks for Fast Credit Card Fraud Detection. Proc. of the 2002 IEEE International Conference on Fuzzy Systems.

The FBI, federal bureau of investigation, "Digital Forensics: It's a bull market", [Online]. Available: http://www.fbi.gov/ news/stories/2007/may/ rcfl050707. [Accessed 07 05 2015].

Thiruvadi, S. and Patel, S. (2011). Survey of Data-mining Techniques used in Fraud Detection and Prevention. Information Technology J., 10(4), pp.710-716.

V. Baryamureeba and F. Tushabe. The Enhanced Digital Forensic Investigation Process Model. In Proceedings of the 4th Annual Digital Forensic Research Workshop, Baltimore, MD. Citeseer, 2004.

Veena H Bhat, Prasanth G Rao, Abhilash R V,P Deepa Shenoy,Venugopal K. R. "A Novel Data Generation Approach for Digital Forensic Application in Data Mining", Second International Conference on Machine Learning and Computing, 2010.

Viaene, S., Derrig, R. & Dedene, G. (2004). A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. IEEE Transactions on Knowledge and Data Engineering 16(5): 612-620

Vogt, W. Paul. SAGE Quantitative Research Methods. London: SAGE, 2011. Print.

Von Altrock, C., J. Yen, R. Langari, and L. A. Zadeh (1995) "Fuzzy Logic Applications in Europe", Industrial Applications of Fuzzy Logic and Intelligent Systems, IEEE Pres, Chicago, 275-310.

W. W. Bratton, "Enron and the dark side of shareholder value", [Online].Available: http://papers.ssrn.com/sol3/ papers.cfm? abstract_id=301475. [Accessed 07 05 2015].

Warren, C. and Pearson, M. (2013). A Look Ahead: Supply Chain Forensics.

Weatherford, M.(2002). Mining for Fraud. IEEE Intelligent Systems, July/August, 4-6.

Wheeler, R. & Aitken, S. (2000). Multiple Algorithms for Fraud Detection. Knowledge-Based Systems 13(3): 93-99.

Wikipedia, the free encyclopedia, "Enron scandal", [Online]. Available: http://en.wikipedia.org/wiki/Enronscandal. [Accessed 07 05 2015].

Williams, G. J. and Huang, Z.(1997). 'Mining the Knowledge: Mine the Hot Spots Methodology for Mining Large Real World Databases', 10th Australian Joint Conference on Artificial Intelligence, Published in Lecture Notes in Artificial Intelligence, Springer-Verlag, December, Perth, Western Australia.

Williams, G.(1999). 'Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries', Proceedings of the 3rd Pacific-Asia Conference in Knowledge Discovery and Data Mining, Beijing, China

Willig, Carla. Introducing Qualitative Research In Psychology. Maidenhead, England: McGraw Hill/Open University Press, 2008. Print.

Woolf, Emile, Suresh Tanna, and Karam Singh. Quantitative Analysis. Plymouth [England]: Macdonald & Evans, 1985. Print.

Xu, J., Yu, Y., Chen, Z., Cao, B., Dong, W., Guo, Y. and Cao, J. (2013). MobSafe: cloud computing based forensic analysis for massive mobile applications using data mining. Tsinghua Science and Technology, 18(4), pp.418-427.

Yamanishi, K., Takeuchi, J., Williams, G. & Milne, P. (2004). On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. Data Mining and Knowledge Discovery 8: 275-300.

Zheng R, Li J, Chen H, Huang Z., "A framework for authorship identification of online messages: writing-style features and classification techniques". Journal of the American Society for Information Science and Technology, February; 57(3), pp.378– 93, 2006.

Zheng R, Qin Y, Huang Z, Chen H., "Authorship analysis in cybercrime investigation", In: Proc. 1st NSF/NIJ symposium. ISI Springer-Verlag; pp. 59–73, 2003.

Zoysa Kasun and Yasantha Hettirarchi (2008), Analyzing Huge data sets in Forensic Investigations.

APPENDIX A: Questionnaire and Interview Question

Forensic Investigation Survey

0% complete

Page 1: Page 1

1 What is your gender? * *Required*

- Male
- Female

2 Select the continent you reside * *Required*

- North America
- South America
- Asia
- Africa
- Australia
- Europe
- Antarctica

3 What is the name of your company or university? * *Required*

4 What sector does your company or your university educational area fall into? * *Required*

- Education
- Oil and Gas
- Banking
- Consultancy
- Automobile
- Other

a If you selected Other, please specify:

5 Do you directly involved in any forensic related investigations? * *Required*

- Yes Full-time
- Yes Part-time
- No

6 Can you explain in more details how you carry out your responsibilities? * *Required*

7 What specific investigative technique or tool do you utilize? ✱ *Required*

- [ ] AFFLib incorporating Advanced Forensic Format (AFF)
- [ ] WEKA
- [ ] ProDiscover IR
- [ ] AutoMiner
- [ ] Other

a If you selected Other, please specify:

8 Who are your targets when using forensic tools? ✱ *Required*

9 How will you rate these investigative tools you use? ✱ *Required*

Please don't select more than 1 answer(s) per row.

Please select at least 1 answer(s).

Having trouble with the format of this question? View in tableless mode

	Excellent	Good	Neither good or bad	Poor	Very poor
Visualization of data	☐	☐	☐	☐	☐
Speed	☐	☐	☐	☐	☐
Easy of us	☐	☐	☐	☐	☐
Help System	☐	☐	☐	☐	☐

10 Do these tools have any drawbacks? ✱ *Required*

- ○ No
- ○ Other

a If you selected Other, please specify:

11 Can you recommend some advance forensic tools that you are aware of that can make your work better but are not currently being used? ✱ *Required*

12 Are you aware of current advancement in forensic investigation study and research, if so, can you share some with me? ✱ *Required*

13 What forensic framework do you typically follow (e.g. DFWR)? ✱ *Required*