

A Unified Forensics Analysis Approach to Digital Investigation

Ali Alshumrani^{1,2}, Nathan Clark^{1,3} and Bogdan Ghita¹

¹ Centre for Cyber Security, Communications and Network Research (CSCAN), University of Plymouth, UK

² Department of Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia

³ Security Research Institute, Edith Cowan University, Western Australia

ali.alshumrani@plymouth.ac.uk

n.clarke@plymouth.ac.uk

bogdan.ghita@plymouth.ac.uk

Abstract: Digital forensics is now essential in addressing cybercrime and cyber-enabled crime but potentially it can have a role in almost every other type of crime. Given technology's continuous development and prevalence, the widespread adoption of technologies among society and the subsequent digital footprints that exist, the analysis of these technologies can help support investigations. The abundance of interconnected technologies and telecommunication platforms has significantly changed the nature of digital evidence. Subsequently, the nature and characteristics of digital forensic cases involve an enormous volume of data heterogeneity, scattered across multiple evidence sources, technologies, applications, and services. It is indisputable that the outspread and connections between existing technologies have raised the need to integrate, harmonise, unify and correlate evidence across data sources in an automated fashion. Unfortunately, the current state of the art in digital forensics leads to siloed approaches focussed upon specific technologies or support of a particular part of digital investigation. Due to this shortcoming, the digital investigator examines each data source independently, trawls through interconnected data across various sources, and often has to conduct data correlation manually, thus restricting the digital investigator's ability to answer high-level questions in a timely manner with a low cognitive load. Therefore, this research paper investigates the limitations of the current state of the art in the digital forensics discipline and categorises common investigation crimes with the necessary corresponding digital analyses to define the characteristics of the next-generation approach. Based on these observations, it discusses the future capabilities of the next-generation unified forensics analysis tool (U-FAT), with a workflow example that illustrates data unification, correlation and visualisation processes within the proposed method.

Keywords: Data Correlation, Data Heterogeneity, Digital Forensics, Digital Forensics Tools.

1. Introduction

The continuous development of technologies, network services and communications alongside the growing number of digital devices adopted and utilised by communities has enabled enormous data expansion. The adoption of modern technologies has also been accompanied by an increase in associated criminal activities, which recorded more than 5.1 billion data breaches in 2021 (Luke Irwin, 2022), presenting several challenges to digital forensics (Chabot, et al., 2015; Chikul, Bahsi and Maennel, 2021). With more than 90% of criminal activities leaving a digital footprint, digital forensics is arguably critical to deal with these crimes (Cellebrite, 2021). Evidence can now be extracted across many devices and delivered in various data formats such as structured, semi-structured and unstructured data (Reeve, 2013). However, the complexity of enormous data volumes and diverse data types, fragmented across different levels and spanning various digital sources, has changed the nature of digital evidence. The proliferation of file system platforms, smartphones, cloud applications, and IoT sensors has led to a constant increase in the volume of data and data heterogeneity at a high rate (Aggarwal and Davis, 2018). Unfortunately, to date, current methodologies and tools tend to be siloed around specific technologies or services. These isolated approaches to examination and analysis introduce significant challenges when aiming to cross-correlate evidence. For that reason, investigating multi-source complex digital information is challenging and requires substantial cognitive load and intensive manual labour (Adderley and Peterson, 2020).

The existing approaches and tools are not widely researched/developed with features to support multi-source correlation analysis. Instead, they primarily rely on experts' knowledge and cognitive ability to map the relationships between diverse data points (Mohammed, Clark and Li, 2018). It is evident that there is a crucial need to develop a robust, unified, automated method to process, integrate, analyse and correlate disparate data within one unified tool. One potential approach to discovering such activity is to cross-correlate evidence across data sources, reveal comprehensive data relationships, and identify inconsistent data or missing parts of evidence. The development of this approach could help to identify a complete set of evidence, reveal hidden data, detect inconsistencies, and identify anti-forensic activities in a timely manner. Therefore, this paper

discusses the future capabilities of a next-generation forensic analysis tool by presenting its ability to harmonise, correlate and visualise disparate data automatically, in a cognitively simple way, and in a forensically and timely fashion.

The remainder of this paper is structured as follows. Section 2 thoroughly evaluates the capabilities and limitations of current digital forensic tools. Section 3 discusses the current state of the art. Section 4 presents a novel approach to unifying digital forensics. Section 5 discusses the conclusions and directions of future research.

2. Evaluation of Digital Forensics Tools

A wide range of digital forensics tools and applications have been developed with features that aid digital forensics practitioners in acquiring, examining, analysing, and reporting digital evidence. This section provides an analysis of the current capabilities and limitations of forensic tools. The tools have been carefully selected based on several criteria, such as their diversity of features and functionalities, recency, and popularity among digital investigation practitioners. Finally, the chosen software is categorised and comprehensively evaluated based on their capabilities of the technical parameters associated with core digital forensics categories, as shown in Table 1.

Table 1: A comprehensive evaluation of digital forensics tools capabilities

Digital Forensics Tools		Parameters	Acquisition & Examination Phase							Analytics Phase					
			Supported Data Source							Limited Data Sources			Automating Multisource Data Analysis		
			File System	Network	Smartphone	Social Media	IoT	Database	Cloud	Automated Analysis Supported	Basic Correlation Analytics	Basic Visualization	Anti-Forensics Detection	Advanced Correlation	Advanced Visualization
File system	Autopsy	✓	✗	✓	✗	✗	✗	L	✗	P&L	P&L	✓	✗	✗	✗
	Detego Unified Forensics	✓	✗	✓	✓	✓	✓	N/A	✓	P&L	✓	✓	✗	✗	N/A
	EnCase	✓	✗	✓	✓	✓	✓	N/A	✓	P&L	✗	✓	✗	✗	✗
	Magnet Forensics AXIOM	✓	✗	✓	✓	✓	✓	N/A	✓	P&L	P&L	✓	✗	✗	N/A
	Nuix Investigate	✓	✗	✓	✓	✓	✗	✓	✓	P&L	P&L	✓	✗	✗	N/A
	OS Forensics	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
	Forensic Explorer	✓	✗	✗	✗	✗	✗	✗	✗	L	✗	✓	✗	✗	✗
	FTK	✓	✗	✓	✓	✓	N/A	✓	✓	P&L	P&L	✓	✗	✗	✗
	X-Ways Forensics	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
Network	NetDetector Suite	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	L	✗	✗	✗
	Network Miner	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	L	✗	✗	✗
	Wireshark	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	L	✗	✗	✗
	Xplico	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	L	✗	✗	✗
Smartphone	Belkasoft Evidence Center	✓	✗	✓	✓	✗	✓	✓	✓	✗	✗	✓	✗	✗	N/A
	Cellebrite PathFinder	✓	✗	✓	✓	✓	✓	N/A	✓	P&L	P&L	✓	✗	✗	✗
	Mobiledit Forensic Express	✗	✗	✓	✓	✗	✗	N/A	✓	P&L	✗	✓	✗	✗	✗
	Oxygen Forensic Detective	✓	✗	✓	✓	✓	✓	✓	✓	P&L	P&L	✓	✗	✗	N/A
	Paraben E3 Universal	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	N/A
	XAMN Analyse	✗	✗	✓	✓	✓	N/A	N/A	N/A	✗	✗	✓	✗	✗	✗
Database	Sanderson Forensics	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
	SysTools SQL Log Analyzer	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗

(✓): Addressed, (✗): Not Addressed, (L): Limited Features, (P): Partially Automated, (N/A): Not Applicable.

It can be seen from Table 1 that these tools have been designed to operate within the confines of particular technologies, whether that be devices, such as computers and smartphones, or types of data, such as network data and databases. A wide range of tools exist, but many of them were developed for file-systems based evidence, for instance, Autopsy, EnCase, Nuix, FTK, and X-Ways forensic tools, with abilities to some extent to deal with other platforms, such as smartphones and internet applications. With a limited capability to perform advanced data analytics, they are limited to conducting data correlation within one single data source or basic data visualisation analysis. With regard to network investigation, a few network data analysis tools exist, such as Wireshark, NetworkMiner and Xplico, but with limited examination features to investigate network logs. Indeed, the network tools evidently lack incorporating any other types of evidence, as they were primarily

developed to monitor and troubleshoot network logs. Whereas smartphones tools, such as Belkasoft, Cellebrite, Modiledit, Oxygen, Paraben, and XAMN, were developed with a great feature to deal with digital resources apart from smartphones, such as file systems, web applications, IoT and cloud data. However, their functionalities mainly reside in fundamental investigation techniques, such as data acquisition and examination phases. Only partial attention is given to developing an advanced investigation method, such as evidence automation concerning multiple data sources.

The following points provide further detailed analysis of the capabilities of these tools against the current challenges in the digital forensics discipline:

- **Evidence Volume:** These investigation tools struggle to process and identify potential evidence among a large volume of data involving multiple evidence resources in a reasonable amount of time (Martinez-Mosquera, Navarrete and Lujan-Mora, 2020). Several experiments have been conducted to evaluate the efficiency of these tools, which show that several hours were taken to perform the acquisition and examination of data evidence (Horsman, 2019). For instance, an experiment aimed to identify items of interest using data signature analysis of data evidence contains 10 million files, and it concluded that tools such as Encase and Nuix required a long time to only identify potential data (Quick and Choo, 2018a). Technically, these experiments indicate that these tools perform poorly against the growing volume of data.
- **Data Heterogeneity:** The current tools are incapable of integrating a large volume of data heterogeneity; even though some of these tools have the ability, to some extent, to extract complex data, such as unstructured data, they still cannot address the issue of providing a complete understanding of evidence data without cognitive overload (Chabot, et al., 2015). Arguably, these tools do not support features that would integrate complex data of different formats automatically (Nordvik, Toolan and Axelsson, 2019). Moreover, the absence of data automation leads to the overloading of digital investigators when they attempt to handle and investigate complex data heterogeneity. Impressively, none of the tools evaluated could provide an investigator with the ability to query evidence across heterogeneous data in a unified and timely manner.
- **Data Correlation:** The developers of Autopsy, Detego, AXIOM, Nuix, and Oxygen state that these tools have data correlation capabilities. Interestingly, these tools can only perform partial data correlation within the same data sources. For instance, the features of the correlation engine module in Autopsy are limited to a few data correlation attributes, such as domain names, email addresses, and phone numbers. Moreover, it is not fully automated and requires forensic investigators to understand the logic of the case before the commencement of the data correlation process and, ultimately, to validate the results. Although the Magnet AXIOM tool utilises a built-in algorithm to determine and tag potential images with illegal content, such as weapons and drugs (Magnet, 2022), the investigator's intervention is still necessary to view and verify the results before analysing the content. Finally, the correlation analysis is limited solely to data file system metadata, with no attention given to recognising different metadata structures from other data sources.
- **Data Visualisation:** Conducting data visualisation analysis can enable examiners to visualise and uncover potential evidence effectively. The selected tools are limited in terms of data visualisation techniques that can function with a reasonable volume of data types. For instance, evidence visualisation in Autopsy is limited to timeline events and fails to combine timeline visualisation with geographical data, although it can extract Global Positioning System (GPS) data from multimedia files and applications. While AccessData reported that FTK can visualise data objects in different views in the same context, its design lacks an interactive navigation system, making it complex for an inexperienced investigator to use (AccessData, 2022). Moreover, the current approaches cannot visualise high-level system events along with the associated low-level traces in a user-friendly interface (Adderley and Peterson, 2020).

It is worth noting that these tools incorporate a wide range of forensic analysis techniques, such as facial recognition, explicit image detection, social network analysis, and data visualisation. Nevertheless, the primary usage of these tools remains in the early stages of the digital investigation process, as their strength resides in the data acquisition and examination phases. Furthermore, the evaluation also concludes that the selected tools cannot support an investigation of a high volume of heterogeneous data in a usable manner. Another limitation is the inability to perform advanced data analysis. For instance, none of these tools can correlate network traffic with another evidence source, such as file systems, or cross-correlate CCTV data with government databases to find data matches. Finally, none of these tools can perform advanced link analysis across all evidence resources within one user interface. Thus, the digital forensics field still lacks sufficient features to conduct a digital investigation in a forensically sound and timely manner.

3. Analysis of the Current State of the Art

To provide an appreciation of the current state of the art, an analysis was undertaken of academic publications related to data volume and heterogeneity, evidence correlation, and data visualisation analytics. Researchers have applied several methods, such as data ontology, machine learning and data graph, to prioritise potential artefacts, remove irrelevant and duplicate data, and correlate and visualise evidence.

Data ontology is beneficial for logically presenting data and relationships between classes of a specific domain. Brady, Overill and Keppens (2015) proposed a Digital Evidence Semantic Ontology (DESO) to organise and categorise relevant artefacts. DESO creates a repository of known artefact attributes, such as data type, location and reference class, to aid examiners in prospectively understanding data and selecting the relevant evidence. The experiment indicated that the DESO could only identify and represent data that remains in its identifiable data structure but fails to process unstructured data. Similarly, Chabot, et al. (2015) introduced Semantic Analysis of Digital Forensic Cases (SADFC). SADFC represents the ontological nature of corresponding data and uses SPARQL data query language to query evidence. During the evaluation, a real malware dataset was utilised, indicating that SADFC could extract related knowledge, perform data analysis, and execute data queries. However, it lacks the ability to automatically analyse and validate cases with a large volume of heterogeneous data. In the same context, Turnbull and Randhawa (2015) proposed Parallax Forensics (ParFor) framework that uses derive system events, such as login information and user events, including browser history and email events, and utilises them to reconstruct social network activities based data ontology mapping. However, this approach is limited to examining the abstraction of users' events and their attributes, whilst no attention has been given to low-level data, such as files and disks. Another framework that seeks to reduce evidence size is proposed by Quick and Choo (2018b). First, the approach seeks to create a logical evidence container and add an intelligence value to disparate evidence by identifying related entities via open-source information, such as social networking sites. The authors suggested using Open-Source Intelligence (OSINT) and developed Digital Forensic Intelligence (DFINT). The researchers utilised a real dataset to evaluate the proposed framework, which confirmed the possibility of semi-automatically scanning and combining data from disparate datasets across different data objects to some extent. However, further research is needed to enhance the system functionalities to reduce the investigation time.

Further ontological studies conducted by Amato, et al. (2020) and Amato, et al. (2019) aimed to employ a semantic-based methodology and Neuro-linguistic Programming (NLP) in their system to integrate and correlate evidence. First, this system utilises Resource Description Framework (RDF) assertions over the associated ontology to accomplish diversified data collection generated by investigation tools. It then uses its rule engine to integrate and correlate evidence data, and finally, a SPARQL module develops and evaluates users' queries against the SPARQL query engine. During the experiment, a collection of logging file events was used, which indicated its ability to detect potential events. Still, it functions to a limited type of data and requires manual intervention to help annotate events. With the same objectives, Chikul, Bahsi and Maennel (2021) developed a ForensicFlow system based on a semantic web to allow knowledge integration and semantically query data relationships. This ontology consists of data sources, data extraction modules, data analysis layers, data knowledge layers, and finally, event-based and forensic artefacts that aim to standardise and define different data types. While evaluating the capabilities of the proposed approach, the researchers simulated a ransomware dataset containing file system data and memory dump and applied the proposed system to the simulated data. The experiment result stated that the system is valuable in revealing and querying evidence data from one data source, but it fails to perform data validation.

Different studies sought to identify data correlation based on metadata analysis. Raghavan and Raghavan (2013) implemented the AssocGEN engine to analyse and determine associations among evidence data. The experiment result indicated that AssocGEN could not support all metadata attributes of file systems applications, making it unsuitable for dealing with the varieties of file systems data that existed today. It also indicates that the data correlation aspect still needs improvement to automate the identification of corresponding data without requiring investigator interventions. Likewise, Mohammed, Clark and Li, (2018), introduced a similar concept for determining artefact associations by developing an algorithm to merge different datasets through data characterisation and harmonisation. The characterisation process identifies the nature of the data through the metadata, and the harmonisation process combines the data. The evaluation shows that this algorithm performs optimally in some respects but cannot integrate all binary data fields, making the harmonisation process less accurate. Finally, these two approaches are limited to some data types, making them less robust and generalised for addressing current crime categories.

Noel, et al. (2016) proposed an alternative approach named CyGraph based on the Neo4j data graph technique. The CyGraph aims to unify, correlate, and visualise network events. First, the system collects and stores network events, security events, and the associated vulnerabilities. It then utilises its built-in model to predict possible attack paths and critical vulnerabilities across cyber entities. To some extent, this approach can aid examiners in executing data queries and expressing graph patterns visually. However, it only covers the predefined data in the data model and does not allow the investigator to manipulate data relationships interactively. In a further approach, Aggarwal and Davis (2018) attempted to standardise and integrate evidence data by converting relational and RDF models into data graphs by leveraging the (Bouhali and Laurent (2015) method. The approach first transforms RDF models by exporting databases into a CSV file format to allow the user to select the schema and transform it into a Neo4j data graph. Then, the Neo4j browser is used to graphically explore and interact with the transformed schemas. However, this approach is incapable of capturing schema conflicts, and it also fails to identify data subset relationships.

In a further effort, Okolica (2017) developed the Temporal Event Abstraction and Reconstruction (TEAR) tool, and employed a machine learning algorithm to aid in confirming the identification of data patterns. Technically, this algorithm initiates a hierarchy of system events and applies a pattern-matching technique to highlight a high-level event with low-level activities. Similarly, Schelkoph, Peterson and Okolica (2019) leveraged the advantages of TEAR and incorporated it to develop the Property Graph Event Reconstruction (PGER) system. The PGER system aims to normalise and identify data correlation by adapting a native graph property of a zero-index traversal to help discover adjacent nodes of related system events, such as web history and download data. However, both TEAR and PGER methods need to be evaluated on a larger scale of data types and a greater data volume. Later, Adderley and Peterson (2020) leveraged the PGER system and introduced the Temporal Analysis Integration Management Application (TAIMA). TAIMA aims to visualise graphically and reconstruct evidence events based on a timeline analysis mechanism. First, TAIMA recognises timestamp events as a research parameter within file system data attributes, such as event logs, registry logs, and link files. Then, it displays the identified high-level events with their associations on a single screen. The researchers utilised a simulated case study to evaluate the effectiveness of TAIMA; the results highlighted that TAIMA could conduct timeline analysis and filter the number of system events based on timeline analysis. However, its functions are limited to timeline analysis and only recognise executable files and web history data. Finally, the analysis presented above is summarised in Table 2.

Table 2: Summary of the literature review

N	Author(s)	Context	Objective(s)	Method	System Name	Mode	Limitation
1	(Brady et al., 2015)	DV	Categorisation of potential data.	Ontology	DESO	Case Study	<ul style="list-style-type: none"> Partially automated. Applicability in complex data.
2	(Chabot et al., 2015)	DV & DH	Automatically represent evidence events.	Ontology	SADFC	Simulated	<ul style="list-style-type: none"> Lacks to validate data automatically. Scalability.
3	(Chikul et al., 2021)	DH & DC	Automatically extract and reconstruct disparate data.	Ontology	Forensic Flow	Simulated	<ul style="list-style-type: none"> Limited to specific data types. Lacks to validate data queries. Scalability.
4	(Turnbull and Randhawa, 2015)	DH	Provide a unified representation of multiple data sources.	Ontology	ParFor	Case Study	<ul style="list-style-type: none"> Data validation. Limitation in data correlation. Applicability in real data.
5	(Amato et al., 2020, 2019)	DH & DC	Integrate & correlate evidence data.	Ontology & NLP	NA	Simulated	<ul style="list-style-type: none"> Partially automated. Limitation in data types.
6	(Quick and Choo, 2018)	DV & DH	Identify the related entity.	Framework	DFINT & OSIT	Real Data	<ul style="list-style-type: none"> Limitation in data types. Investigation time.
7	(Raghavan and Raghavan, 2013)	DH & DC	Identify data association	Engine	AssocGE-N	Real Data	<ul style="list-style-type: none"> Limitation in data types. Partially automated.
8	(Mohammed et al., 2018)	DH	Data characterisation & harmonisation.	Algorithm	NA	Real Data	<ul style="list-style-type: none"> Limited data sources. Lack to merge all binary data.
9	(Noel et al., 2016)	DC & DVIs	Unification model to correlate and visualise network events.	Data Graph	CyGraph	Case Study	<ul style="list-style-type: none"> Applicability in complex data. Scalability & usability.
10	(Aggarwal and Davis, 2018)	DH	Transform relational and RDF models into data graph properties.	Data Graph	NA	Case Study	<ul style="list-style-type: none"> Lacks to capture all data. Applicability in real data.
11	(Okolica, 2017)	DH	Correlate evidence data.	Data Graph & ML	TEAR	Real Data	<ul style="list-style-type: none"> Limited data types. Applicability in complex data.
12	(Schelkoph et al., 2019)	DC	Normalise and correlate evidence data.	Data Graph & ML	PGER	Real Data	<ul style="list-style-type: none"> Limited data types. Applicability in complex data.
13	(Adderley and Peterson, 2020)	DC & DVIs	Graphically correlate & visualise evidence data.	Data Graph	TAIMA	Case Study	<ul style="list-style-type: none"> Limited evidence events. Timeline analysis only. Applicability in real data.

DV: Data Volume, DH: Data Heterogeneity, DC: Data Correlation, and DVIs: Data Visualisation

Researchers have contributed to overcoming challenges in digital forensics by proposing several methods and techniques based on technologies such as data ontology, machine learning, and data graphs. Unfortunately, most of these methods attempt to articulate a specific obstacle or focus on a particular data type. Thus, many limitations remain, indicating the need for further research to fill the identified gaps, enhance investigation productivity, and overcome existing challenges. For instance, a comprehensive research should include data unification and data automatically to merge disparate data and provide usability to ensure that examiners can use the solutions in a way that requires little effort in identifying and correlating potential evidence across various data within one graphical interface. Therefore, this study emphasises the necessity of establishing a unified forensics analysis solution to integrate and unify the ubiquitous nature of data heterogeneity and identify the relationships between all digital events and user activities across different sources of evidence.

4. Next-Generation U-FAT

To facilitate a comprehensive understanding of the functionalities and features required of a next-generation *Unified Forensics Analysis Tool* (U-FAT), a list of common crimes was compiled, categorised and linked with the necessary corresponding digital analyses to address investigation questions and meet investigation objectives. Having defined these core requirements, U-FAT is proposed and discussed.

4.1 Crime Classification

Current criminal investigation cases have come to be involved in interconnected data from across different technologies, requiring advanced digital forensics analysis techniques that must function alongside each other to finalise an investigation in a timely fashion. Figure 1 summarises some of the most common crimes, associating them with examples of digital evidence sources. It also highlights some types of forensic analysis that could be employed to conduct criminal investigations and address the investigation objectives.

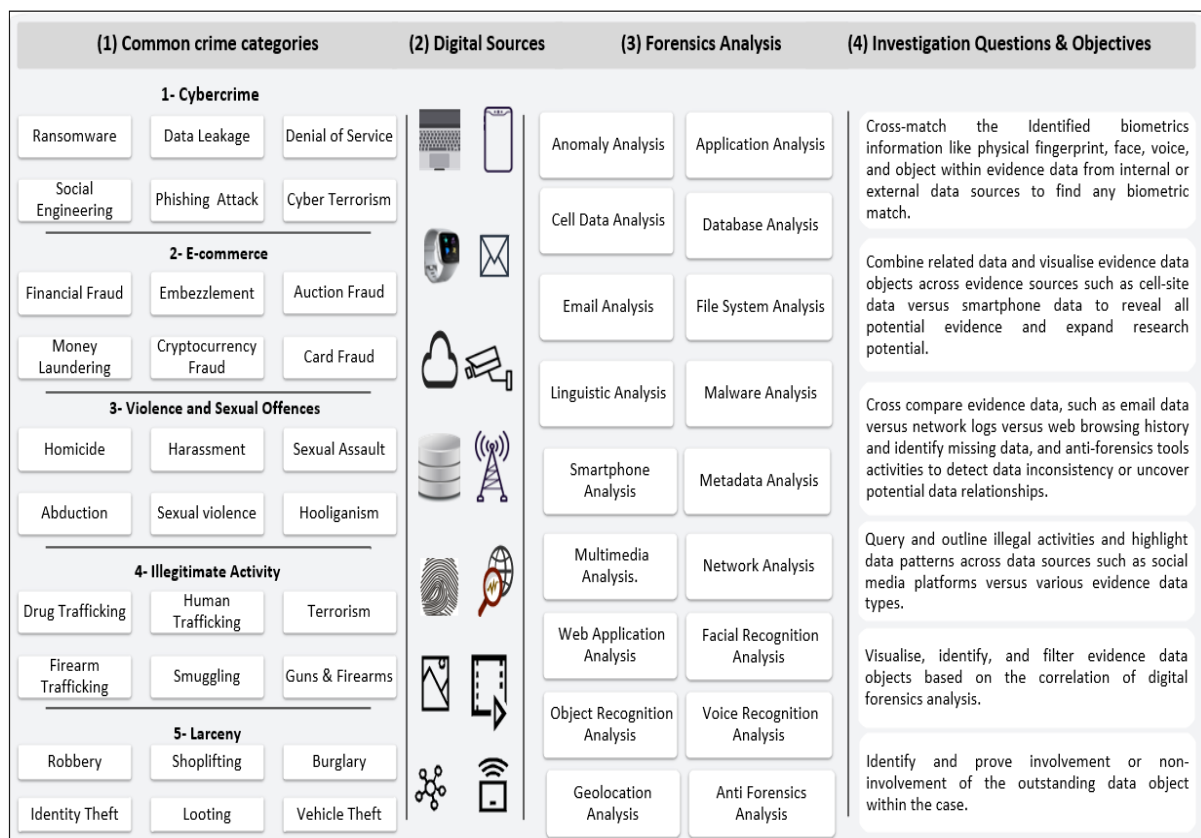


Figure 1: Classification of crimes and forensics analyses

It is worth noting that in Figure 1, each crime might involve a variety of data sources containing different data formats, which in turn require several types of forensic analysis to conduct a proper digital investigation. Two examples are provided below as an illustration:

- **Cybercrime:** Investigating cyber hacktivist activities, such as those involving ransomware, denial-of-service attacks, social engineering, and online scams, can be challenging, as these crimes involve multiple connected computers, cloud system resources, applications, and network devices. Therefore, several forensic examinations are required based on each data source, for example: file system, network, malware, database, and application analyses.
- **Violence, illegitimate activity, and larceny crimes:** Although these types of offences might not be categorised as digital crimes, still, digital data can be employed to support an investigation. Evidence could come from digital devices, such as laptops and smartphones, or external data, including public internet applications, CCTV, and call data records. Whilst investigating such data, digital investigators will perform different forensic analyses, such as: application, multimedia, linguistic, biometric, social network, cell data records, databases, and social network assessments to identify potential evidence and correlate the results to establish appropriate answers.

In summary, Figure 1, highlighted the complexity of current digital investigation cases, the limitations of digital tool capabilities outlined in Table 1, and the existing gaps in digital forensics methods pinpointed in Table 2 have formed the main drive underpinning in this research.

4.2 System Workflow

The proposed approach comprises a unified system capable of forensically performing end-to-end the investigation of disparate data in a time-efficient, cognitively simple, and conclusive manner. A holistic workflow of the proposed U-FAT system is illustrated in Figure 2. It shows various featured functions to perform data analytics side by side, for instance: mapping evidence components, detecting anomalies and data inconsistencies, revealing data query matching, discovering data patterns, and permitting cross-platform analysis of complex, interconnected data within a single unified graphical interface.

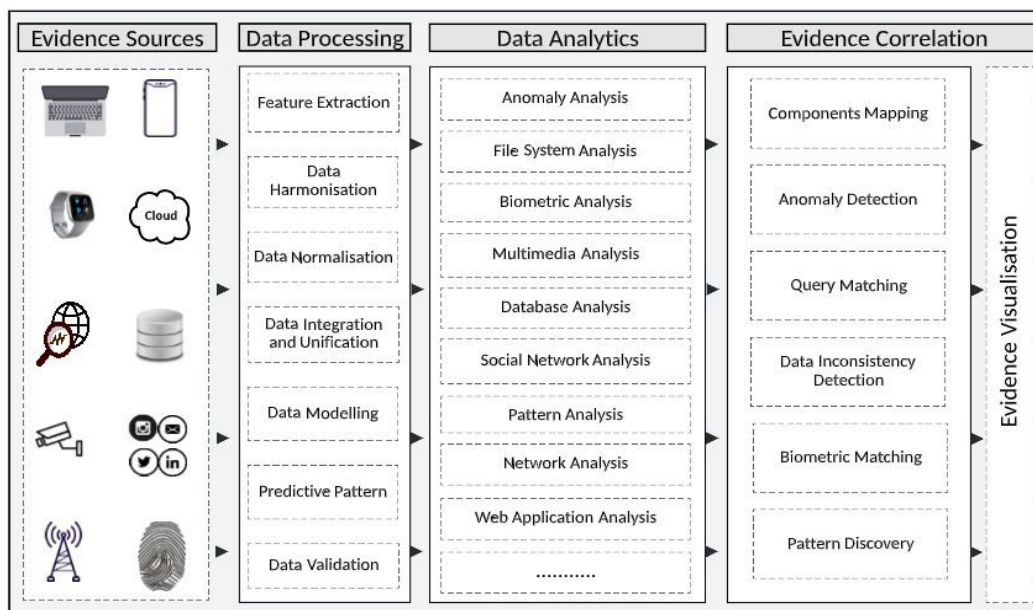


Figure 2: Holistic illustration of the U-FAT process

The initial functionalities of the system aim to define and extract disparate data depending on different data filtering and extraction mechanisms. Then, various data examination techniques are applied in the pre-processing stage to eliminate duplicate or irrelevant data, identify potential data and pass it to the data processing phase. In the data processing phase, the system applies data normalisation, harmonisation, and unification techniques, to interpret complex data into a standard and appropriate normalised data format, grouping and unifying them based on their attributes and characteristics. Afterwards, different evidence analytics methods will be undertaken to detect data anomalies, patterns, biometrics, and many more based on the requirements of the investigation.

To technically illustrate the proposed approach, Figure 3 presents a workflow of the data unification and correlation processes in a case involving the need to analyse evidence data, such as: file system, email, network logs, web applications, and social media, as this could be the case in many investigation crimes.

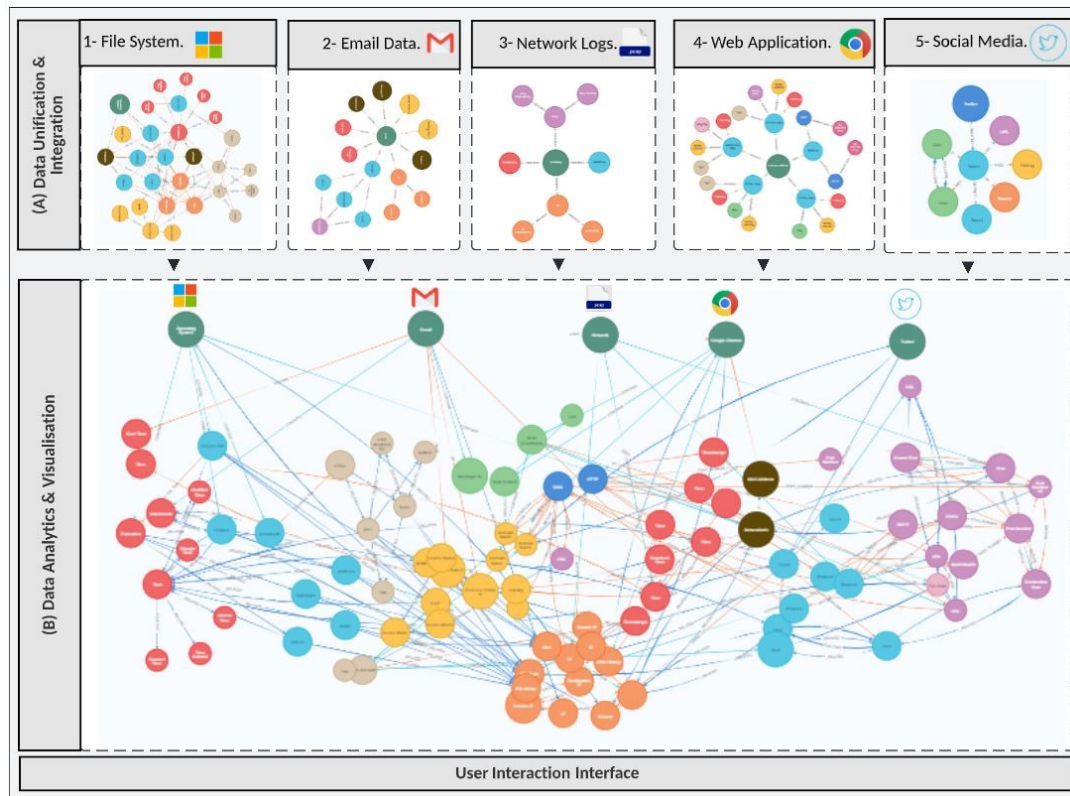


Figure 3: Workflow example of data unification, correlation, and visualisation in Neo4j data model.

The main benefit of the data unification technique is to enable advanced forensic analysis across different data platforms. Figure 3 illustrates the core functional components of the U-FAT, where the initial process aims to identify different data sources, extract and categorise data based on each data attribute, and then utilise its built-in mechanisms to unify and correlate data in a unified view of complete evidence. Figure 4 presents a technical example of how these processes have enabled advanced forensic analysis.




Data Sources		Evidence Extraction, Normalisation & Harmonisation & Data Analytics							Automated Evidence Correlation Analysis								
<div> Network Logs</div>	Time	Source IP	Destination IP	Protocol	Source	Destination		Time	Source IP	Destination IP	Protocol	Source Port	Destination Port	Application	Correlation Status	Result	
	06/07/2020	54.240.48.52	20.90.153.243	HTTPS	443	65361		06/07/2020 11:22:39	54.240.48.52	20.90.153.243	HTTPS	443	65361	Chrome	Mismatched	Anti-forensic Detected	
	06/07/2020	54.240.48.52	52.108.89.13	HTTPS	443	26560		06/07/2020 11:23:16	54.240.48.52	52.108.89.13	HTTPS	443	26560	Chrome	Mismatched	Anti-forensic Detected	
	06/07/2020	54.240.48.52	157.56.236.134	SMTPS	578	50094		06/07/2020 15:21:37	54.240.48.52	157.56.236.134	SMTPS	578	50094	Gmail	Matched	Successfully Correlated	
	06/07/2020	54.240.48.52	157.56.236.134	SMTP	578	50094		06/07/2020 15:22:52	54.240.48.52	157.56.236.134	SMTP	578	50094	Gmail	Mismatched	Anti-forensic Detected	
	06/07/2020	54.240.48.52	52.108.89.13	HTTPS	443	26567		06/07/2020 15:26:42	54.240.48.52	52.108.89.13	HTTPS	443	26567	Microsoft Edge	Mismatched	Anti-forensic Detected	
	06/07/2020	54.240.48.52	157.56.236.134	SMTPS	578	50094		06/07/2020 15:29:17	54.240.48.52	157.56.236.134	SMTPS	578	50094	Outlook	Matched	Successfully Correlated	
<div> Email Data</div>	Email Analysis							06/07/2020 16:21:37	54.240.48.52	157.56.236.134	SMTPS	578	50094	Gmail	Matched	Successfully Correlated	
	Time	MessageId	From	To	Subject	Protocol	SPF	Application									
	06/07/2020 15:21:37	wpqo1rue	Email1@gmail.com	E1@gmail.com	Hi	SMTPs	pass with IP 54.240.48.52	Gmail									
<div> Web Application</div>	06/07/2020 15:22:52	wpqo1rue	Email1@gmail.com	E1@gmail.com	Hello	SMTPs	pass with IP 54.240.48.52	Gmail									
	06/07/2020 15:29:17	wpqo1rue	Email1@gmail.com	E1@gmail.com	Hey	SMTP	pass with IP 54.240.48.52	Outlook									
	06/07/2020 16:45:11	wpqo1rue	Email1@gmail.com	E1@gmail.com	update	SMTPs	pass with IP 54.240.48.52	Gmail									
	Web Application (Browser-Web History) Analysis							06/07/2020 15:29:17	54.240.48.52	157.56.236.134	SMTPS	578	50094	Outlook	Matched	Successfully Correlated	
	Date Accessed	URL	Referrer URL	Title	Domain	Program Name											
	Null	Null	Null	Null	Null	Chrome											
	Null	Null	Null	Null	Null	Microsoft Edge		06/07/2012 16:21:37	54.240.48.52	157.56.236.134	SMTPS	578	50094	Gmail	Matched	Successfully Correlated	

Figure 4: Workflow example of data harmonisation and correlation processes.

In this example, the proposed approach utilised metadata attributes of evidence data, including network logs and email data from applications such as Outlook or Gmail, and historical data of web applications such as Google Chrome or Microsoft Edge, to automatically harmonise and unify in a standardised format. This approach would enable advanced data analytics to function harmoniously against harmonised data automatically; for instance, mapping artefacts against a single timeline, data correlation analysis and detecting the presence of anti-forensics techniques. Figure 4 indicates that it is possible to identify when a suspect might have been using data-hiding techniques, such as private web browsing or deleting emails from the client machine. Web histories from a suspect machine can be cross-correlated with network traffic to identify them, and the client email records can be compared against network and server-side records, therefore, contributing to addressing the current challenges in digital forensics and reducing and eliminating the investigation time and manual cross-correlation.

In the U-FAT approach, various analytics modules are integrated and function side by side to perform advanced evidence analytics. The analytics automation process provides the capability to harmonise, query, correlate and analyse evidence across evidence data. Notably, the correlation analytics stage is considered the cornerstone of the U-FAT approach, enabling digital investigators to work interactively to query, review and edit data relationships across a complete evidence case. Having the ability to retrieve, correlate and graphically visualise potential data within one tool could provide more insight into the output of the analytics process. Empowering forensic investigators with the proposed approach can considerably shorten the investigation time and reduce the cognitive for more significant insights to be gained through different visualisation techniques in respect of heterogeneous data with different views.

5. Conclusion

Unifying heterogeneous evidence and incorporating digital forensics analysis has become increasingly important in digital forensics investigations. Current approaches offer a range of forensics analyses and semi-automated data techniques. However, the methods discussed have yet to be able to process, integrate and correlate a vast amount of data heterogeneity in an automated and harmonised fashion. Therefore, it is essential to integrate and unify multiple data sources and process advanced analytics in one intuitive interface to overcome the limitations and complexity of the current frameworks of evidence investigation. Furthermore, an intelligent query feature is proposed to support the flexibility of enabling data interrogation functions across technologies and services. Combining these features with an automated mechanism to understand, correlate and visualise data intelligently will aid digital investigators in addressing questions in a more timely and cognitively efficient manner. Future research will focus on developing a common forensic information interchange language to permit cross-platform disparate case analysis and on developing advanced forensic analysis capability.

References

- AccessData, 2022. Zero in on evidence faster—recognized around the world as the standard in computer forensics software, [online]. Available at: <<https://accessdata.com/assets/pdfs/LIT-FTK-7.4.2.pdf>> [Accessed 10 November 2022].
- Adderley, N. and Peterson, G., 2020. Interactive temporal digital forensic event analysis. In: *IFIP Advances in information and communication technology*. pp. 39–55. https://doi.org/10.1007/978-3-030-56223-6_3
- Aggarwal, D. and Davis, K. C., 2018. Employing graph databases as a standardization model for addressing heterogeneity and integration. In: Rubin, S.H., Bouabana-Tebibel, T. eds., *Advances in intelligent systems and computing*. Springer International Publishing, Cham, pp. 109–138. https://doi.org/10.1007/978-3-319-56157-8_6
- Amato, F., Castiglione, A., Cozzolino, G. and Narducci, F., 2020. A semantic-based methodology for digital forensics analysis. *Journal of Parallel and Distributed Computing*, 138, pp.172–177. <https://doi.org/10.1016/j.jpdc.2019.12.017>
- Amato, F., Cozzolino, G., Moscato, V. and Moscato, F., 2019. Analyse digital forensic evidences through a semantic-based methodology and NLP techniques. *Future Generation Computer Systems*, 98, pp.297–307. <https://doi.org/10.1016/j.future.2019.02.040>
- Bouhali, R. and Laurent, A., 2015. Exploiting RDF open data using NoSQL graph databases. In: *IFIP Advances in Information and Communication Technology*. New York: Springer. pp. 177–190. https://doi.org/10.1007/978-3-319-23868-5_13
- Brady, O., Overill, R. and Keppens, J., 2015. DESO: Addressing volume and variety in large-scale criminal cases. *Digital Investigation*, 15, pp.72–82. <https://doi.org/10.1016/j.diin.2015.10.002>
- Cellebrite, 2021. Digital Intelligence Benchmark Report, [online]. Available at: <<https://cellebrite.com/en/digital-intelligence-benchmark-report-2021>> [Accessed 10 November 2022].
- Chabot, Y., Bertaux, A., Nicolle, C. and Kechadi, T., 2015. An ontology-based approach for the reconstruction and analysis of digital incidents timelines. *Digital Investigation*, 15, pp.83–100. <https://doi.org/10.1016/j.diin.2015.07.005>

- Chikul, P., Bahsi, H. and Maennel, O., 2021. An ontology engineering case study for advanced digital forensic analysis. In: *Lecture notes in computer science* (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 67–74. https://doi.org/10.1007/978-3-030-78428-7_6
- Horsman, G., 2019. Tool testing and reliability issues in the field of digital forensics. *Digital Investigation*, 28, pp.163–175. <https://doi.org/10.1016/j.diin.2019.01.009>
- Luke Irwin, 2022. Free Infographic: List of data breaches in 2021 [online]. IT Gov. Eur. Blog. Available at: <https://www.itgovernance.co.uk/infographics/list-of-data-breaches-in-2021> [Accessed 11 January 2022].
- Magnet, 2022. Magnet IEF - Artifact-First Investigation [online]. Available at: <https://www.magnetforensics.com/products/magnet-ief/> [Accessed 22 September 2022].
- Martinez-Mosquera, D., Navarrete, R. and Lujan-Mora, S., 2020. Modeling and management big data in databases—a systematic literature review. *Sustainability*, 12, p.634. <https://doi.org/10.3390/su12020634>
- Mohammed, H., Clarke, N. and Li, F., 2018. Automating the harmonisation of heterogeneous data in digital forensics. In: *European conference on information warfare and security, ECCWS*. pp. 299–306.
- Noel, S., Harley, E., Tam, K. H., Limiero, M. and Share, M., 2016. CyGraph: Graph-based analytics and visualization for cybersecurity. In: *Handbook of statistics*. Elsevier B.V., pp. 117–167. <https://doi.org/10.1016/bs.host.2016.07.001>
- Nordvik, R., Toolan, F. and Axelsson, S., 2019. Using the object ID index as an investigative approach for NTFS file systems. *Digital Investigation*, 28, pp.S30–S39. <https://doi.org/10.1016/j.diin.2019.01.013>
- Okolica, J.S., 2017. Temporal event abstraction and reconstruction. Air Force Institute of Technology Wright-Patterson AFB United States
- Quick, D. and Choo, K.-K.R., 2018a. IoT device forensics and data reduction. *IEEE Access* 6, pp.47566–47574. <https://doi.org/10.1109/ACCESS.2018.2867466>
- Quick, D. and Choo, K.-K.R., 2018b. Digital forensic intelligence: Data subsets and Open Source Intelligence (DFINT+OSINT): A timely and cohesive mix. *Future Generation Computing Systems*, 78, pp.558–567. <https://doi.org/10.1016/j.future.2016.12.032>
- Raghavan, Sriram and Raghavan, S. V., 2013. AssocGEN: Engine for analyzing metadata based associations in digital evidence. In: *2013 8th International workshop on systematic approaches to digital forensics engineering (SADFE)*. IEEE, pp. 1–8. <https://doi.org/10.1109/SADFE.2013.6911541>
- Reeve, A., 2013. *Managing data in motion*. Elsevier. <https://doi.org/10.1016/C2011-0-07758-X>
- Schelkoph, D. J., Peterson, G.L. and Okolica, J.S., 2019. Digital forensics event graph reconstruction. In: *Lecture notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*. Springer Verlag, pp. 185–203. https://doi.org/10.1007/978-3-030-05487-8_10
- Turnbull, B. and Randhawa, S., 2015. Automated event and social network extraction from digital evidence sources with ontological mapping. *Digital Investigation*, 13, pp.94–106. <https://doi.org/10.1016/j.diin.2015.04.004>