Digital Forensics 2.0: an automated, efficient, and privacy preserving digital forensic investigation framework

By

Robin Verma

Under the Supervision of

Dr. Gaurav Gupta, and
Dr. Donghoon Chang

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Digital Forensics 2.0: an automated, efficient, and privacy preserving digital forensic investigation framework

By

Robin Verma

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

to

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

MAY 2018

Abstract

The current state of digital forensic investigation is continuously challenged by the rapid technological changes, the increase in the use of digital devices (both the heterogeneity and the count), and the sheer volume of data that these devices could contain. Although data privacy protection is not a performance measure, however, preventing privacy violations during the Digital Forensic Investigation, is also a big challenge. The investigator gets full access to the forensic image including suspect's private data which may be sensitive at times as well as entirely unrelated to the given case under investigation. With a perception that privacy preservation and the completeness of investigation are incompatible with each other, the digital forensics researchers have provided solutions to address the above-stated challenges that either focus on the effectiveness of the investigation process or the data privacy preservation. However, a comprehensive approach that preserves data privacy by neither affecting the capabilities of the investigator nor the overall efficiency of the investigation process, is still an open problem. In the current work, the authors have proposed a digital forensic framework that uses case information, case profile data and expert knowledge for automation of the digital forensic analysis process; utilizes machine learning for finding most relevant pieces of evidence; and preserves data privacy in such a way that the overall efficiency of the digital forensic investigation process increases without affecting the integrity and admissibility of the evidence. The framework improves validation to enhance transparency in the investigation process. The framework also uses a secure logging mechanism to capture investigation steps to achieve a higher level of accountability. Since the proposed framework introduces notable enhancements to the current investigative practices more like the next version of Digital Forensics, the authors named it 'Digital Forensics 2.0', or DF 2.0 in short.

Keywords: Digital Forensics, Digital Forensic Investigation, Digital Forensic Process Model, Survey, Data Privacy, Digital Forensic Framework, Automation, Machine learning, Classification, Clustering.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

The modern computing devices which receive, store and process information in digital form have permeated every way of human life everywhere across the globe. If we would like to visualize all day to day actions in an individual's life who uses the present-day digital technology in one way or another, one can think of it like a virtual web connecting the individual to her surroundings and other people she interacts with.

Every event which has some involvement of digital technology, represents one strand of this imaginary web. For example, an event may include any of the following – an individual traveling from one place to another with her geographical coordinates getting recorded on the smartphone, or she is communicating over an email, or she is making a grocery purchase using her bank debit card, or she is engaging in some activity on any of the contemporary online social networks. A strand connects a person to her environment or with another individual or a group of people while possessing some digital information, which got created or exchanged during the process.

All individuals keep on adding these strands to their virtual web using the digital devices they carry, or wear, or interact with while doing their daily activities. Every device generates or records some form of digital data for the respective activity (a strand in the imaginary web). The data associated with a strand can either be stored on the devices that the individuals use, and/ or on devices that are present in the vicinity of where these individuals are, and/ or on an entirely different storage space far away from the geographical locations of the individuals.

Overall, the structure of an individual's virtual web could be very complex at times depending upon the number of digital devices and the number of persons involved, their respective geographical locations, the variety, volatility and volume of the digital data created or exchanged, and the medium of exchange.

Digital Forensics is the field of study that aims to reconstruct the sequence of events in case a security incident is reported, and hence find the suspected person(s) responsible for the same. In view of the above-stated visualization of the current digital landscape in the form of a virtual web, one can get an abstract outline of the challenges faced by digital forensic professionals while reconstructing the sequence of events in a given case.

Any criminal activity in today's world would inadvertently have some digital involvement [54]. When a criminal (or a malicious) activity is reported, the law enforcement personnel has to examine the digital devices and the data involved in the case to find out or indicate the person responsible. On an abstract level, a digital investigation aims to produce all possible evidence that can be used in a court of law to explain what could have happened in a given case [17].

The digital forensics investigation process is a collection of sub-procedures that include identification, collection, analysis, reporting, and presentation of the digital evidence [48]. The identification refers to recognizing the potential source of evidence from the crime scene. The collection relates to a careful acquisition of digital data so that the integrity of the evidence can be ensured. The analysis is required to find out links which connect digital data to events; that might be further related to a suspect. After establishing a relationship between the data and the suspect, a report is generated to document the same formally. Eventually, the report is presented in a court of law that may prove or disprove the allegations against the person under trial.

1.1 Motivation

The digital forensic science has evolved a lot since the first Digital Forensic Research Workshop [48]. However, there are still some research problems that are continuously challenging the researchers and practitioners to date.

The first and foremost challenge is the ever-growing data storage capacity of digital devices [55]. The large volume of data increases the time requirements for the data acquisition and the data analysis processes [41]. Moreover, since the number of cases that involve digital evidence in some form is on the rise all over the world, the digital forensic investigators are facing a pressing need for reducing the investigation time per case [3].

The second challenge is thrown by the increasing diversity of digital devices that are becoming available in the market [31]. A digital forensic personnel has to continuously strive for finding new ways (through software as well as hardware means) to acquire and analyze such devices [35]. The software diversity deals with a massive number of file-types, ever-evolving Operating Systems, the newly developed innovative applications, and other software advancements concerning contemporary digital devices. On the hardware front, diversity of sensors, chips, circuit modules and other hardware units that produce unique data streams presents a challenge for digital forensics. Although providing a solution to both of the above-stated diversity challenges takes only a one-time effort for the practitioners and researchers, however, the rate at which these parameters change keeps them on their toes.

Furthermore, people tend to use separate devices for communication, entertainment and productivity purposes. Hence the number of individuals who own and use more than one digital devices at a time is increasing [24]. Another study by Facebook in 2016 reveals that 94% teens in France and 98% teens in Germany own multiple devices [25]. The Pew Research Center published a report in 2015 stating that around 36% of US adults own all three devices, namely a smartphone, a computer, and a tablet [5]. Another survey by Pew in January 2017 has revealed that 77% of US adult population owns a smartphone, 78% owns a desktop or laptop, and 51% owns a tablet computer [50]. Although the survey presents separate figures for the three devices, one can safely assume that individuals who own multiple devices are a significant part of the US population today. The people in other regions of the world either share similar trends or would achieve the same figures shortly. The rise in the number of devices owned per person would increase the average number of exhibits seized in a new case, thus increasing the respective investigation time and efforts.

Even after finding their ways to acquire and analyze the new digital devices, the digital forensic examiners face the third challenge from the rapidly changing technological advancements that change the rules of the game now and then [28]. The technological progress that poses a challenge to investigators is concerned with the increasing list of devices that are going digital every day, thanks to the new software and hardware innovations. The devices in everyday use which get equipped with computational, communication and digital storage capability, commonly referred to as the Internet of Things (IoT), pose new investigative challenges to the digital forensic process [47]. Any investigation involving such devices would require knowledge about how the data is produced, stored and communicated to these devices.

The fourth challenge, which is not directly connected to the functioning of the digital forensic investigation, is data privacy protection during the digital forensic investigation [4]. The Digital forensic investigators always get full access to the contents of seized storage media which according to them is necessary for achieving completeness. Apart from containing potential evidence files, the seized storage media also contains owner's private data, which may be sensitive at times like private/family pictures and videos, business-related digital documents, medical diagnostic or treatment reports, commercial software with license information, and much more. Investigator's open access to these private files is a threat to owner's data privacy [72].

The data privacy protection is also related to need for transparency in the digital forensic investigation that ensures only case-relevant data are accessed from the seized media and remaining private files are not affected [22]. There is a pressing need for finding means to fix accountability of the investigator in case a data privacy breach happens during the investigation. The two sister agencies that work in close collaboration with digital forensic personnel, namely the Police and the regular forensic laboratories, are facing difficulties related to transparency and accountability. The case of Annie Dookhan is a good example of the same [23]. To the best of authors' knowledge, there are no reported instances of professional misconduct against digital forensic investigators till date, however, it is high time that the community should adopt self-regulatory ways to improve the transparency as well as the accountability of the digital investigation process.

The current thesis started with a premise that data privacy during the digital forensic investigation is an issue for the three parties; which are involved in a digital forensic investigation process, namely the general public (representing the potential victim and the suspect), the digital forensic investigator (representing the investigate agency), and the cyber lawyers (who debate on the legal outcome of a case in court of law).

The first part of the work aims to find out the ground truth about the possibility of data privacy violations in the digital forensic investigation process. Hence, the author circulated three surveys for the same, one for each of these three classes of participants. The analysis of the responses shows a lack of professional ethics among some of the investigators, lack of legal support to protect data privacy for lawyers, and, confusion among the general public regarding their data privacy rights.

Digital forensic frameworks to date have focused on addressing the above-stated challenges either in separation or well-defined scenarios with controlled environmental conditions. The author believes that the problem of data privacy in digital forensic investigation process cannot be addressed in isolation, and hence privacy preservation should be a part of the digital forensic investigation model. The data privacy protection should be incorporated in such a way that the incorporation does not have any effect on the efficiency of the investigator, or her investigative powers.

The second part of this thesis proposes a novel digital forensic model that uses automation to increase the efficiency of the investigative process while keeping a tab on data privacy violations during the process.

Digital forensic researchers have introduced automation in the digital forensic investigation to simplify and speed up the overall process. However, the automatic identification of potential pieces of evidence during the analysis phase is still an open problem.

The third part of this thesis discusses implementation and analysis of Machine Learning algorithms for predicting the potential evidence in a given case. The author has also used Machine Learning techniques to find the privacy quotient of the files in the same case. Both of these details, the evidential score, and the privacy value are presented to the investigator so that her job is simplified while the data privacy of files that have no evidential value and are highly private is also protected at the same time.

1.2 Contributions

The main contributions of the present thesis are as follows:

A. The thesis proves the need for privacy preservation during digital forensic investigation process.

- The author conducted a study containing three different surveys, one for each stakeholder in a digital forensic investigation, namely the investigators, the cyber lawyers and the general public, that aim to capture respective perceptions of data privacy during the investigation process.

- The responses show a lack of professional ethics among some of the investigators, lack of legal support to protect data privacy for lawyers, and, confusion among the general public regarding their data privacy rights.

- The findings indicate towards the pressing need for a privacy-preserving digital forensic investigation framework that protects data privacy without compromising on investigator's efficiency and performance during the digital forensic investigation.

- The surveys focus on the Indian context of the digital forensic investigation and hence all survey participants are from India. Although the surveys were circulated in India, the concerns raised and the results obtained are relevant for the global population as well.

B. Proposing a novel next generation of Digital Forensic Investigation Model that can incorporate automation, efficiency as well as data privacy protection.

The author proposes a new digital forensic framework that incorporates forensic image preprocessing, tool-independent automation, machine learning based filtration of most relevant evidence and their privacy level evaluation. The framework proposes a new way in which the state of the art digital forensic research and systems could be combined in one place to realize the following.

- Increased investigative efficiency by saving in the investigation time and efforts
- Improved investigative accuracy by using multiple tools at the same time
- Better investigative planning via automation
- Improved validation
- Data privacy protection for forensically non-relevant private files
- Enhanced transparency and accountability
- Building expert knowledge for forensic investigation, education, training, and multi-agency collaborations

C. Implementation of Machine Learning (ML) techniques to assist in identification of potential evidence files while also evaluating their privacy level simultaneously, which saves the investigator's time and helps her to avoid accidental data privacy violations.

- The author has modeled the evidence identification as a two-class classification problem (a type of supervised learning technique).

- The prototype implementation obtained reasonably good results; by minimizing the false negatives using the 'Bagging' technique.

- The privacy evaluation of files is modeled as a clustering problem (an unsupervised learning technique).

- The clustering which has been implemented using the K-means ML algorithm also obtained reasonably good results.

1.3 Thesis organization

The author presents a novel digital forensic framework, DF 2.0, that uses automation to improve the efficiency of the digital forensic investigation process; and utilizes machine learning to predict potential pieces of evidence while avoiding intentional as well as accidental data privacy violations from the hands of the digital forensic investigator. The contents of the thesis are organized in the following way:

- Chapter 2 provides an overview of digital forensics, data privacy preservation in digital forensic investigation models, and the next generation of digital forensic implementations. All these topics are closely related to the current thesis work.

- Chapter 3 presents a study containing three different surveys, one for each stakeholder in a digital forensic investigation, namely the investigators, the cyber lawyers and the general public, which aim to capture respective perceptions of data privacy during the investigation process. The findings of the three surveys point towards the pressing need for a privacy-preserving digital forensic investigation framework that protects data privacy without compromising on investigator's efficiency and performance during the digital forensic investigation.

- Chapter 4 provides details about the novel digital forensic framework that uses case information, case profile data and expert knowledge for automation of the digital forensic analysis process; utilizes machine learning for finding most relevant pieces of evidence; and preserves data privacy in such a way that the overall efficiency of the digital forensic investigation process increases without affecting the integrity and admissibility of the evidence. The framework also strengthens validation which enhances the overall transparency of operations in the investigation process. The framework uses a secure logging mechanism which records the investigation steps to achieve a higher level of accountability. Since the proposed framework introduces significant enhancements to the current investigative practices more like the next version of Digital Forensics, the authors have named it 'Digital Forensics 2.0', or DF 2.0 in short.

- Chapter 5 discusses the implementation of machine learning algorithms and techniques on the digital forensic artifacts seized during the digital forensic investigation. The author have used two class classification algorithms to train machine learning models which predict whether a given file is a potential evidence for the case under investigation or not. Moreover, the results of the machine learning models are further improved by reducing the false negatives with the help of bagging technique. The authors have also used clustering algorithms to estimate the privacy quotient of all files, which helps the investigator to avoid intentional or unintentional data privacy violation of the files, that have a high privacy level, but no relative evidential value.

Chapter 2

BACKGROUND

2.1 Digital Forensics

Personal Computers brought a kind of revolution in the lives of ordinary people, as they started using computers in their daily life activities. Most of the people used the power of PC to enhance their quality of life and welfare of the society; whereas there were some who tried to exploit these capabilities for their malicious gain. The investigative agencies who were given the job of examining such cases of computer abuse, found out pioneering ways of collecting evidence from them. The first mention of the use of computers in crime, is in the book titled ‘Crime by Computer’ by Parker and Parker [49]. Pollitt [54] gives a sequential and most detailed history of how computer forensics started and evolved with time.

2.1.1 A brief history

Researchers and practitioners of this field started working on standardizing the procedures that are used to carry out an investigation on devices and entities that contain digital data. Pollitt [51], in the year 1995, introduced a four-step model that included Acquisition, Identification, Evaluation, and Admission as evidence. These steps ensure admissibility of the collected evidence in the court of law. In the year 2000, Noblett et al. [46] suggested a three-level hierarchical model that gave directions about the investigation, the organizational protocols and the procedures that need to be followed in a computer forensics case, which has to meet the same standards as those of a typical forensics investigation.

The year 2001 will be remembered as a turning point in the field of digital forensics when a consensus document (Palmer [48], also known as the Roadmap Document) was formulated during the first Digital Forensic Research Workshop, that gave a formal definition to the field of 'Digital Forensics'. The Model proposed in this document had 6 steps namely, Identification, Preservation, Collection, Examination, Analysis, and Presentation. Reith et al. [56] extended these steps in the year 2002 and came up with a total of 9 steps which according to them were more close to the traditional forensic model, which finds an easy acceptance in the courts. They advocated for a preparation and strategy planning, before going for the preservation step; and concluded the investigation with a new step of returning the evidence.

The next year, the year of 2003, saw four research works that presented their own versions of digital forensics models. Carrier and Spafford [13] started with their Integrated Digital Investigation Process (IDIP), which mapped the digital investigative process to the physical investigative process with the help of 17 phases that were further organized under 5 groups. Stephenson [68] published the second paper of 2003, where he proposed an End-to-End Digital Investigation Process which had nine steps that were built on top of the Roadmap Document. In the third paper of 2003, Carrier [11] talks about putting layers of abstraction at every step of his defined digital forensics investigation process, which makes the job of verifying the 'credibility of evidence' easy. The last paper of 2003, by Mocas [42] introduces context to the investigation model. According to the author, every context would enforce certain conditions on the investigation process and hence would influence the output of the investigation.

The year 2004 witnessed four more papers that focused on the digital forensic models. The first among them was by Baryamureeba and Tushabe [8], who suggested improvements to the IDIP, and segregated the crime scene into two; one focuses on what happened in the computer and the second concentrates on the physical objects. This segregation according to the authors was a better approach to remove inconsistencies from the investigation. The second paper that followed was from Beebe and Clark [9], which proposed to divide the forensic analysis step into subtasks that work in coherence to achieve the investigative goal in an objective based goal-oriented manner. The third research was put forward by Carrier and Spafford [14] who conceptualized the digital investigation as a cause-effect scenario, where a change in the state of a digital object is caused by events. The goal of the investigation is to reconstruct the sequence of events. The last paper of 2004, presented by Pollitt [53] took inspiration from an old Indian story where six blind men try to explain how an elephant looks like based on their personal observation by touching specific parts of the elephant. The moral of the story is that every person may have their own explanation of a concept/idea which may be true based on their observation; however, the bigger picture may be entirely different. The author brings the same argument to digital forensic models, by comparing the NIST Incident Response Model, the Roadmap document model, and the Zachman framework [75]. The author suggests, after pointing out similarities within these three, that digital forensics should be considered as a group of sub-tasks that are driven by purpose and bounded by constraints.

The year 2005 paper by Ruibin [64] introduces the concept of expert knowledge which is possessed by the investigator and can be used to carry out the analysis phase in an efficient way. The authors also put forward the notion of ‘case-relevance’ which is defined as the property of a piece of information, which is provided with the case or known to the investigator as a virtue of her knowledge of the trade, that can be used to answer the investigative questions of a given case in a better way.

There are more digital forensic models that have critically reviewed classical models (like the ones discussed above) and improved upon them. These improved models have evolved on the basis of which all new research or implementation challenges within the digital forensic field they promise to solve, and which new hardware and software technological advancements they have incorporated to achieve the same.

Ieong [34] states that after a security incident involving digital devices is reported, the security agencies rush to the place where it happened to find out reasons that lead to the event. They aim to find all plausible details and potential digital objects that could be linked to the person or entity responsible for the episode and prove the same in the court of law. Carrier [12] says that Digital Forensic Investigation (DFI) should ensure a legal backing for the collected digital evidence; and further suggested [16] to supplement the investigation with physical evidence (like DNA matching) to establish a relationship between the suspect/entity and the digital device. Similarly, Cohen [20] also includes attribution as a process in his digital forensic process model (DFPM) which finds a link between the digital evidence and the suspected person.

The forensic investigation approaches that are used by digital forensic personnel have advanced a lot during last few years. All of this progress can be credited to two primary reasons. First is the continuous learning that comes from the experience which is gained while solving real-life cases. The second reason is the endless research efforts that resulted in the invention of new tools and technologies that have changed the way the digital investigation is carried out.

2.1.2 Formal definition

Digital forensics, as it was defined in first DFRWS conference [48], is "the use of scientifically derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation and presentation of digital evidence derived from digital sources for the purposes of facilitating or furthering the reconstruction of events found to be criminal or helping to anticipate the unauthorized actions shown to be disruptive to planned operations". This definition is considered to be the most inclusive definition of digital forensics [38].

Willassen and Mjolsnes [74] gave a new definition as "the practice of scientifically derived and proven technical methods and tools towards the after–the-fact digital information derived from digital sources for the purpose of facilitating or furthering the reconstruction of events as forensic evidence". The difference between the above-stated definitions by Palmer [48] and Willassen and Mjolsnes [74] is that the later removed the word 'criminal'. Willassen's definition broadens the scope of digital forensics to include corporate and other type of investigations, that need not be criminal in nature.

Pollitt [53] defines digital forensics "as a group of tasks, steps or sub-processes followed during the investigation process making the whole process more flexible in choosing methods and technologies towards its goals”.

Kohn [38] has presented a refined definition of digital forensics as “a specific, predefined and accepted process applied to data stored digitally or digital media using scientifically proven and derived methods, based on a solid legal foundation, to extract after-the-fact digital evidence with the goal of deriving the set of events or actions indicating a possible root cause, where reconstruction of possible events can be used to validate the scientifically derived conclusions”.

Kohn gave this definition after critically analyzing six Digital Forensic Process Models, namely Lee et al. [40], Casey [17], Carrier and Spafford [14], Baryamureeba and Tushabe [8], Ciardhuain [19] and Cohen [20]. The comparative analysis of these DFPMs (as it is given in Kohn [38]) is presented in table 2.1. It may also be noted that Kohn’s definition, when compared to that of Willassen’s definition, expands the characteristics of the extraction process and potential forensic tools.

2.1.3 Digital forensics branches

Other forms of Digital Forensics namely, live forensics, proactive forensics, and network forensics are named that way because they target a specific hardware or a special infrastructure ([10], [60], [44]). Live forensics aims for evidence collection from running systems, especially targeting the RAM as it holds valuable information that may be lost afterwards [2]. Proactive forensics aims to actively gather potential evidence that may help in stopping a security incident before it happens, or maximize the chances that the guilty would be caught with minimum investigation efforts and cost [44].

A new way of investigation called the computer forensic triage, was proposed by Rogers et al. [60] which aims to extract relevant evidence by examining user profile from her computer usage, internet activity and other subsequent actions within a time interval of interest.

Table 2.1: Comparative summary of the DFPM as discussed in Kohn [38].

Phase	**Process**	**Lee**	**Casey**	**Carrier & Spafford**	**Baryamureeba**	**Ciardhuáin**	**Cohen**
Preparation	Policy and Procedure						
	Infrastructure Readiness			S-P	S-P		
	Operational Readiness			S-P	S-P		
Incident	Detect	P	P	S-P	S-P	P	
	Assess		P				
	Confirm			S-P	S-P		
	Notify			S-P	S-P	P	
	Authorise			S-P	S-P	P	
	Deploy			P	S-P		
	Approach Strategy					P	
	Search					P	
	Recover		P				
	Seize		P				
	Preserve	S-P	P	S-P			
	Transport					P	P
	Store					P	P
DFI*	Collect	S-P	S-P	S-P		P	P
	Authenticate						
	Examine					P	P
	Harvest		P				
	Reduce		P				
	Identify	P	P			P	P
	Classify	S-P	P				
	Organize		S-P				
	Compare	S-P	S-P				
	Hypothesise					P	
	Analyse		P				S-P
	Attribute	P	S-P				S-P
	Evaluate	S-P					
	Interpret	S-P					S-P
	Reconstruct	P		S-P	S-P		S-P
	Communicate						
	Review			P			
Presentation	Present report		P	S-P	S-P	P	P
	Decide						
	Disseminate					P	P

Legend:
*Digital Forensic Investigation.
P = Process.
S-P = Sub-process.

2.2 Data privacy in Digital Forensic Process Models

A digital forensic process model presents the way in which an investigation should proceed from the time of the first response to an incident till the investigation is completed. It acts as a user manual for the investigators, to guide them on how to collect and analyze potential evidence from devices.

Although there are plenty of digital forensic process models (DFPM) discussed in digital forensic literature, however the author has found only one framework (Van Staden [70]) that incorporates privacy of data into the digital forensic investigation of a computer system. Van Staden proposes a framework that protects the privacy of third party during a digital forensic investigation with the help of a profiling and filtering mechanism. Depending on the sensitivity of data being queried, a decision is taken whether the data should be presented to the examiner or not. The paper focuses on enhancing the privacy in multi user environments,that are subjected to post incident investigations.

The author assumes that the third party data is totally unrelated to the suspect whose devices are being investigated. Hence, according to the author the data privacy of the third party, is separate from data privacy of the suspect which may not be the case. The author has also said that a Privacy Enhancing Technology (PET) model would help in the investigative process to ensure that the privacy of the third party is preserved. The PET model accepts queries from investigator that are restricted to text-only or file types. The files that are returned by the forensic tool, as a reply to the queries, are checked for the similarity or dissimilarity (named as the *difference*) between their owners. There is no mention of accuracy of this mechanism in terms of false positives and false negatives. It is worthy to note here that false positives may be acceptable in digital forensics investigations, even though they would increase the investigative effort. However, the false negatives could allow potential evidence to slip away, hence can change the directions of the investigation and alter the course of justice.

Sometimes, during the investigation it may happen that the investigator might not have a clear idea about exactly what type of files she is looking for. In such a situation, the current system's expectation of receiving focused queries may actually result in repetitive privacy breach warnings to the investigator. The author has not addressed this situation in their paper. The paper offers a new direction to the field of digital forensics, but a lot of open questions remain un-answered.

So, the above research gap motivated the author of this thesis to think of designing a new digital forensic framework that includes data privacy protection as a core feature. The introduction of data privacy protection should not have any impact on either the investigative powers of the examiner or the efficiency of the investigation process. The thesis presents DF 2.0, a novel digital forensic framework in chapter 4 that ensures that the above-stated requirements are met.

2.2.1 Data privacy solutions under controlled environment

There are some excellent papers that have provided solutions to the data privacy protection problem in the digital forensic scenario. However, their solutions are either designed for a specific environment and not generic in nature; or the privacy protection works as a separate module that has performance implications. Some of the notable papers are mentioned below.

Dehghantanha and Franke [22] have defined the same as a cross-disciplinary field of research and named it as 'privacy-respecting digital investigation'. They also talk about the present challenges and opportunities that the field has to offer.

Aminnezhad et al. [4] state that digital forensic investigators face a dilemma whether they should protect suspects' data privacy or achieve completeness in their investigation. The paper also states that there is a lack of awareness among professional digital forensic investigators regarding suspects' data privacy, which could result in an unintentional abuse. There have been attempts to protect data privacy during digital forensic investigation using cryptographic mechanisms. Law et al. [39] have proposed a way to protect the data privacy using encryption. The authors talk of encrypting data set on an email server and indexing the case related keywords, both at the same time. The investigator gives keyword input to the server owner, who has the encryption keys, to get back the emails that contain the keyword.

Hou et al. [33] propose a mechanism to protect the privacy of data on third party service provider's storage center form the investigator using homomorphic and commutative encryption. At the same time, the mechanism also ensures that the service provider does not get to know the queries that were fired by the investigator. Hou et al. [32] talk of a similar solution on a remote server.

Shebaro et al. [67] use Identity Based Encryption to carry out a network traffic data investigation in privacy preserving setting. Gou et al. [30] put forward generic privacy policies for network forensic investigations.

Croft et al. [21] have proposed a mechanism where data is compartmentalized into layers of sensitivity, less private data on lower layers and highly private data on higher layers. Investigator's access to private information is controlled by initially restricting his access to the lower layers first. The investigator is required to prove his knowledge of the low-level layers, to get access to higher level information.

The Df 2.0 framework ensures that the data privacy protection is incorporated into the digital forensic model and hence does not have any impact on the efficiency of the investigation process.

2.3 Next generation of digital forensics

The author of this thesis got an opportunity to design a digital forensic framework that incorporates data privacy protection into the design. The author of this thesis used this opportunity to incorporate efficiency-enhancing measures (like preprocessing 4.3, automated-analysis 4.4, and Machine Learning implementation for potential evidence prediction 4.5) that take the performance of the framework to next level. The thesis chapters 4 and 5 describe the same in more details.

Here are some notable research works that proposed the next level of digital forensics. They either incorporated high levels of hardware performance or advocated the use of automation as a performance enhancement measure; or both.

Ayers [6] describes the limitation of the first generation of digital forensic tools that are struggling with the huge volumes of data involved in modern day investigations. The author proposes several parameters to measure efficiency, together with the requirements that need to be incorporated into the second generation of digital forensic tools. The author also proposed processing architecture of second generation tools which utilizes Beowulf clusters, supercomputers, distributed systems, and grid computing. The evidence storage, workflow management and software reliability of the second generation tools are also discussed. The paper provides requirements and high-level characteristics of the system that was under development.

Garfinkel [28] also talks about the requirement for data standardization and modular mechanisms in the field for digital forensics and digital forensic research.

Van Baar et. al [69] have brilliantly moved the digital forensic processing on a cloud where high-end machines could speed up processing and help different actors involved in a digital forensic investigation to collaborate on a particular case.

Carrier [15] proposed a way to automate searches in digital forensic investigations. Richard et. al [57] suggested a way to handle large-scale digital investigations with the use of distributed computing. They proposed the use of a cluster of distributed computers to facilitate processing and store the images and results at a central data store. The authors suggested the use of automation by all forensic tools so that they may handle the challenges of tomorrow.

Abbott et. al [1] proposed an automated way to correlate events for digital forensic investigation. The authors also demonstrate implementation using publically available digital forensic scenarios and data.

2.4 Summary

There has been an extensive work in the digital forensic field where researchers and practitioners have separately suggested models and frameworks that focus on standardizing operations, boost performance with high-end computing, provide data-privacy on demand, and recommending automation. Machine Learning for potential evidence prediction is still an open problem. The thesis work takes motivation from the above-stated literature and proposes a next version of digital forensics which is explained in the chapter 4.

Next...

One of the initial hurdles that the author of this thesis faced was that there was no previous study that presented the ground truth about the current state of data privacy in the digital forensic investigations. So, the author started his work with three surveys, one for each the digital forensic investigators, the cyber lawyers, and the general public, which were aimed to collect the current state of privacy-preservation during the digital forensic investigation process.

Chapter 3

SURVEY: PERCEPTION OF DATA PRIVACY DURING DIGITAL FORENSIC INVESTIGATION IN INDIA

3.1 Introduction

Privacy is a very complex term to define, as one can have diverse definitions of privacy depending on the context. Privacy can be considered as a tool that enables an individual to control access to her personal space [43]. An individual's personal space in the digital world consists of her data in the form of files. These personal files are either stored on digital devices, else on some local or online storage space.

Digital forensics investigation aims to find all pieces of evidence that link a malicious activity carried out on a digital device to the person responsible. Digital forensic investigators always get full access to contents of the seized storage media, that according to them, is necessary to find all possible pieces of evidence that could help in solving the concerned case. Apart from containing potential evidence files, seized storage media also contain owner's private data like personal or family photographs, videos, business plans, emails, medical documents, financial details, music, movies, commercial software, just to name a few. Investigator's unrestricted access to case unrelated files including owner's private files is a significant threat to owner's data privacy. Secondly, the investigators stop their investigation only after collecting sufficient number of case relevant evidence. There are no well-defined standards or guidelines that help the investigator to decide whether the gathered potential pieces of evidence are sufficient to solve a particular case or not. This lack of clarity motivates the investigators to dig for more, that inevidently increases the chances of data privacy violation.

Legal support is necessary for an accused or a victim to safeguard her data privacy during the investigation of a case and its subsequent trial in a court of law. The cyber lawyers should know all provisions in the law of the land that protect data privacy during a case investigation and following trial. The accused or victim should also be sensitive towards their data privacy rights.

The author wanted to collect the ground truth about the data privacy issues related to an investigation by circulating survey questionnaire among three concerned parties, namely the investigator, the cyber lawyer and the general public. The general public group represents the accused and/or the victim (authors use the word *'suspect'* to address them in rest of the paper) whose storage media are seized for investigation. The surveys focus on the Indian context of the digital forensic investigation and hence all survey participants are from India. Although the surveys were circulated in India, the concerns raised and the results obtained are relevant for the global population as well.

In the first survey [§A.1], the author has taken the views of digital forensic investigators about the privacy of data contained in the forensic image of suspect's storage media that is seized for a particular case. The survey findings show that investigators do not respect the privacy of suspect's data during an investigation. The respondents interestingly accept that they often view suspect's case irrelevant private data and occasionally copy some of it from the case image.

The second survey [§A.2] explores data privacy from cyber lawyer's perspective. The survey findings show that some of the respondents know instances where the suspects have reported misuse of the information gathered during the investigation that was later used to threaten them (the suspects) by the police personnel (responsible for investigation of the given case).

The third survey [§A.3] collects responses from the general public to know their level of awareness about the privacy of data stored on their digital devices. The responses show that people are either confused or unaware of safeguarding their data privacy in case law enforcement agencies seize their digital devices.

An analysis of the literature reveals that this is the first study to collect investigator's, lawyer's and general public's perception of data privacy during the digital forensic investigation. The responses have been collected from all three stakeholders to present an extensive insight into the problem. The surveys have collected responses from India, however the findings put forward profound concerns confronting the global digital forensic community today. The outcomes indicate towards the need of a privacy-preserving digital forensic investigation framework that protects data privacy during the investigation process without compromising either on the completeness of the investigation or the efficiency of investigators. The author proposes a competent digital forensic framework that couples data privacy with completeness while keeping the efficiency of the investigator unaffected.

In rest of the chapter, the words *'investigator'*, *'lawyer'* and *'investigation'* have been used instead of 'digital forensic investigators', 'cyber lawyers', and 'digital forensic investigation' respectively. The first survey with digital forensic investigators as the target audience is referred as the *'investigator survey'*. The second survey with the cyber lawyers as the target audience is referred as the *'lawyer survey'*. The third survey with the general public as the target audience is referred as the *'public survey'*. The words 'private' and 'personal' are used interchangeably in rest of the chapter.

3.2 Survey methodology

Survey research is a well-established field in computer science. Researchers working in digital forensics have also used surveys to understand the viewpoint of the target audience on a specific topic [62, 63]. The survey results help researchers to get a better insight into a particular problem that is very helpful to explore possible solutions.

The first step of survey design in this work included personal interviews with one candidate each for the investigator survey and the lawyer survey. Simultaneously, five potential candidates were also interviewed for the public survey. The answers helped the author to identify a closed set of relevant questions for respective surveys.

The second step of survey design included converting the subjective questions together with their responses to successive objective questions with thorough answer options to choose. The initial questionnaire was shown again to the interviewed candidates to collect their feedback on question formulation. The feedback helped the author to improve the readability, relevance and comprehensiveness of the respective survey questionnaires. The three surveys were then hosted online on survey hosting website ‘surveymonkey.com’.

The investigator survey questionnaire flow is divided into three subsections based on grouping of similar questions:

1. Following the forensic procedure
2. Suitable time to stop the investigation and evidence gathering, and
3. Accessing suspects’ private files

The lawyer survey questionnaire flow is divided into four sub-sections:

1. Minimum number of evidence required
2. Investigation of one case leading to the prosecution of the other
3. Suspects concerned about the privacy of their data, and
4. Misuse and threatening

On the same lines, the flow of questions in the public survey is divided into two sub-sections:

1. Gathering general attitude towards the privacy of data and personally identifiable information, and
2. Checking awareness about digital forensics and the investigation process

The third and final step included sending the online link of the respective surveys to their target audiences. The investigator and the lawyer survey went online in the month of August 2013. The last entry in the investigator survey was received in the month of January 2014, and that for the lawyer survey was received in the month of February 2014. The public survey went online in early September 2013, and the last entry received was in December 2014. The paper includes analysis of the entries received till December 2014 for the three respective surveys. These three surveys cover a holistic picture of data privacy during the process of digital forensic investigation, as the surveys record inputs from all three parties (the investigator, the lawyer, and the general public) involved in the digital forensic case.

3.3 Demographics

The investigator and the lawyer survey includes participants who are experts from the respective fields. All of the participating investigators have undergone professional training and certifications in the field of digital forensics. The participating lawyers are experts on the Indian Information Technology laws and they actively work on cases of cyber crimes and computer frauds in the country.

The investigator survey respondents include digital forensic investigators working on real life cases. They have experience of working in criminal cases as well as corporate investigations. In the investigator survey, a total of 15 digital forensic investigators filled in their responses. The survey received 100% responses from all of the participants with no skipped questions. There is a bigger participation from the private sector, 11 out of 15 respondents are from privately owned digital forensic labs or companies. Rest 4 investigators are working for government forensic labs. 10 out of 15 have a degree in computer science, and the rest 5 are from different background. 7 out of 15 have less than 2 years of working experience in digital forensics field; another 4 are working as investigators for the last 2 to 5 years.The rest 4 have an experience of 5 to 10 years in the field. Table 3.1 shows the number of cases the participants have solved during the course of their career as an investigator.

Table 3.1: Number of digital forensic cases solved as an investigator.

Number of cases Solved	Percent-responses	Actual-responses (out of 15)
Less than 10	40.00%	6
Between 10 to 30	13.33%	2
Between 30 to 50	20.00%	3
Between 50 to 70	13.33%	2
Between 70 to 100	6.67%	1
More than 100	6.67%	1

The lawyer survey respondents are working as cyber lawyers in the state High Courts and the Supreme Court of India. 5 out of 10 participants work with a privately owned firm including one participant who owns a law firm. Another 3 participants work as independent law consultants and the remaining 2 work with government agencies. The lawyer survey also received 100% responses from all of the participants with no skipped questions. The experience of participants as cyber lawyers, in a number of years, is presented in Table 3.2.

Table 3.2: Experience in number of years as cyber lawyer.

Experience (in years)	Percent-responses	Actual-responses (out of 10)
0 to 2	40%	4
3 to 5	20%	2
6 to 8	20%	2
8 to 10	0%	0
10 and above	20%	2

In the public survey 1235 participants filled the complete demographics; 654 people quit before reaching the demographics section. The number of male respondents are 66.6% and rest 33.4% are female. The age-wise categorization of participants is shown in Table 3.3. 17.2% of the respondents have 0-4 years of experience of using computing devices, 21.5% have 4-6 years of experience, and 61.4% of the re-

Table 3.3: Age wise classification of 'public survey' participants.

Age	Number of respondents
Upto 18 yrs.	4.00%
19 - 24 yrs.	61.10%
25 - 34 yrs.	17.80%
35 - 44 yrs.	6.80%
45 and above	10.30%

Table 3.4: Educational qualification of 'public survey' participants.

Educational Qualification	Percentage
High School	5.20
Undergraduate diploma	5.80
Bachelor's degree	56.50
Post-Graduation	29.90
Doctoral degree	2.60

spondents have more than 6 years of experience. The educational status of the participants of public survey are listed in Table 3.4. The demographic of the survey shows that the participants are well educated and have sufficient experience of using computing devices.

The author framed a hypothesis that participant's level of awareness about various privacy issues related to digital documents would be high. This was proved otherwise when all the responses were compiled later.

3.4 Survey 1: privacy from investigator's perspective

The aim of this survey is to assess how digital forensics investigators cater privacy of the accused or victim's data on the seized storage device. Although the number of participants in the investigator survey (digital investigation experts) are limited, yet the collected responses are valuable enough due to the expertise level of these participants in their field. The categorization of the questionnaire is discussed in the following subsections.

3.4.1 Following procedure: Chain of Custody (CoC)

Chain of custody is a legal constraint on people who handle a digital forensic case to track potential pieces of evidence (usually on paper, supplemented with pictures, or sometimes videos), from the time of their seizure till they are presented in a court of law or handed back to the owner after the investigation is over. CoC contains information about the seized exhibit along with the name of personnel, designation and period of custody. CoC is maintained to fix accountability and bring fairness to the overall process.

Author's intention to frame the first two questions around CoC is to check if the investigators are well-versed in the basics of their trade and are serious about their job. When asked about do they follow CoC in cases 14 out of 15 respondents said they fill the CoC form, and the last respondent did not know about it. 11 out of the 14 follow CoC in all of the cases at 'all times', the other 2 follow it 'most of the times but not always' and the last one follows CoC 'only sometimes'.

The next question asks the 14 respondents, who answered with a 'yes' to the previous question, about what encourages them to follow CoC. 2 out of the 14 follow CoC only if the 'case is going to the court of law'. Another 3 follow CoC only after ensuring that the case is important enough. Whereas the rest 9 follow CoC in 'all cases' irrespective of the case going to the court of law, or it is an internal corporate investigation.

The responses to above two questions show a mismatch, where 11 investigators in the first question follow CoC at 'all times', but only 8 among them in the subsequent question say that they follow CoC in 'all cases'. Out of the rest three, 2 follow CoC only if they think the case is important enough, and the last respondent follows CoC only if the case is going to court.

3.4.2 Suitable time to stop the investigation

The first question under this subsection asks investigators whether they stop after finding case relevant potential pieces of evidence or they explore the forensic image further, increasing the chances of encountering suspect's personal files that are irrelevant to the case. 8 out of 15 consider stopping after they have gathered all possible pieces of evidence including both those pertinent to the case as well as those that are totally unrelated. The next 6 out of 15 stop their investigation after they have gathered all possible evidence related to the given case. The last one stops the investigation after collecting a minimum number of evidence needed to prove or disprove a particular case.

The subsequent question asks whether the participant has ever experienced a situation where she gets hold of some pieces of evidence that do not match with the case in hands and could be used in filing a separate fresh case against the suspect. Surprisingly, 7 out of 15 responded with a 'yes, most of the times', another 4 out of 15 responded with a 'yes, only sometimes'. The remaining 4 do not get evidence for new unconnected cases while investigating a given case.

The responses to above-stated questions show that gathering excess of evidence is a regular practice among digital forensic investigators. The habit of searching for more than required either results from investigator's indecision over gathering sufficient pieces of evidence to solve a case or her attempt to gain reputation for discovering distinct evidence that might open a fresh case against the suspect. In each of the situations, such behavior from the investigator opens room for data privacy breach.

3.4.3 Accessing suspect's private files

The first question in this subsection asks participants about their reaction after they encounter the suspect's private files (like personal photographs, videos, songs, business plans, or any form of intellectual property) during the investigation. 6 out of 15 view all such private files, copy some related to the case under investigation as well as others that are not linked to the case but appear illegal or questionable in nature. The other 4 out of 15 view and copy all private files because the files are more likely to be evidence in a given case and other possible cases. The rest 5 out of 15 said that they would view all private files, but copy only those which are related to the case under investigation.

The results show that all participants access suspect's private files that may or may not be associated with the case in hand. Surprisingly, 10 out of 15 would not hesitate to copy suspect's private files if they find any irregularities related to a particular case or otherwise.

Another interesting succeeding question asks whether participants have seen any fellow forensic investigator, in their laboratory or elsewhere, who while investigating a given case copies files like wallpapers, songs, movies, games or commercial software from the case image. 3 out of 15 replied that they have seen their colleagues doing the same in the laboratory they work for themselves. Another 4 have seen investigators copying suspect's non-malicious personal files, but in some other forensic laboratories they visited sometime. One responder replied of not personally seeing anyone copying such files, however she did not see any problem if an investigator did so. The remaining 7 have not seen it happening anywhere and feel that it is not a right thing to do.

Surprisingly half of the participants have seen their fellow investigators in their laboratory or elsewhere who copy suspect's non-malicious content from the forensic images provided for investigation purposes. This unprofessional behavior is a great threat to the data privacy. If the examiner could copy wallpapers, songs, movies, games and application software from suspect's media then security of her private files including personal images, audio-videos and countless types of confidential documents cannot be guaranteed.

The last question before demographics asks whether the participants have seen or heard of any incident where the suspect has reported a misuse of information or the potential pieces of evidence gathered during the case investigation to threaten them. Interestingly, one out of 15 investigators knows about such a reported case. 9 out of 15 replied with a 'no', that they have not heard of any exploitation of information like this across their careers. Rest 5 candidates are not sure if such abuse could even happen.

3.5 Survey 2: privacy from cyber lawyer's perspective

The aim of the lawyer survey is to get insights into a legal aspect of how privacy is catered during a digital forensic investigation and the trial of the respective case in the court of law. The author conducted a pilot interview with one cyber lawyer who is currently working in the *Supreme Court of India* that helped in framing a comprehensive questionnaire for the survey. Although the number of participants in the lawyer survey (cyber law experts) are limited, yet the collected responses are valuable enough due to the expertise level of these participants in their field. The following subsections represent the grouping of the questions.

3.5.1 Completion of a case

The first question asks at what stage, in a case of Cyber Crime and Computer Fraud, the preparation for a trial is complete. 7 out of 10 believe they are ready after they have gathered all possible pieces of evidence both relevant as well as irrelevant to the case. These evidence could be used to prosecute the suspect in a fresh case. 2 out of 10 would stop after they have gathered all pieces of evidence related to the case. The remaining one would stop after collecting a minimum number of potential evidence.

Assuming that the term 'evidence' refers to the number of files that are needed to answer an investigative question or establish a fact during an investigation. The following question asks about the minimum number of evidence that are sufficient to prove or disprove a Cyber Crime and Computer Fraud case in the court of law. 4 out of 10 respondents believe that '1 or 2' evidence are sufficient. 3 participants say '3 to 5' evidence are enough, whereas the remaining 3 think '6 to 10' evidence are required in a given case.

The responses to this question are significant because they set an upper limit on the pieces of evidence needed in a common digital forensic case. At a maximum ten evidence are adequate for digital forensic case, that actually starts with the seizure of digital devices containing hundreds and thousands of files. Except the evidence files, the rest of the data on the seized digital device are irrelevant to the case and may be labeled as private to the suspect.

3.5.2 Investigation of one case leads to prosecution of the other

The first question in this subsection intended to verify the results from investigator survey where the participants get hold of some potential evidence for activities not related to the case in hand that could be used to file a new separate case against the suspect. Asking the same question from cyber lawyers makes sense because all pieces of evidence collected by the team-effort of investigators and lawyers, are compiled and used by cyber lawyers while presenting the case in court of law. 1 out of 10 respondents 'always' gets such a situation where the evidence collected could be used to start a new fresh case against the suspect. 5 candidates experience same situation 'most of the times', while other 3 participants get similar findings 'sometimes'. Only one respondent replied in a 'no'.

3.5.3 Accused or victim asked for legal protection of their data privacy

The three questions in this subsection ask specifically about three privacy supportive laws offered by either the Constitution of India or the Information Technology Act 2000 & 2008 amendment.

The first one asks about number of cases that the participants have handled where the suspect has applied for her 'right of privacy' referring to either the freedom of speech and expression under Article 19(1) (a) or right to life and personal liberty under Article 21 of the Constitution of India, or both. 5 out of 10 lawyers have experienced at least 10 such cases, 3 of them have seen 10 to 30 of such cases. Interestingly, the remaining two participants have observed 30 to 50 and 50 or more such cases each.

The second question asks about the suspect accusing investigative agencies for data privacy breach under section 72A of the (Indian) Information Technology Act, 2000. According to the section, the agencies have been accused of accessing and disclosing suspect's private information, which is irrelevant to the case being investigated. For example, the access and/ or disclosure of personal or family photographs and videos, when the suspect is being investigated for financial fraud. 6 out of 10 answered in a 'yes', with 2 to 5 instances of such cases. The rest 4 replied to in a 'no'.

The third question asks about the suspect accusing investigative agencies for data privacy breach under section 43A of the (Indian) Information Technology Act, 2000. According to the section, the agencies have been accused of improper or negligent handling of suspect's sensitive personal data or information during an investigation of the case. 6 out of 10 answered in a 'yes', with 1 to 5 instances of such cases. The rest 4 replied to in a 'no'.

A similar subsequent question asking participants about number of cases they have solved and others they have knowledge about, where the suspects have requested the court to preserve their private data or files on their seized digital devices. 3 out of 10 have seen up to 10 such cases, while other 2 have seen 10 to 20 of such cases. Interestingly, one of the respondent has knowledge of more than 90 of such cases. The rest 4 out of 10 have not seen such an example till date.

3.5.4 Misuse of information for threatening

The last question before demographics asks participants whether they have heard of any incident where the suspect has reported misuse of information (especially, the collected pieces of evidence after completion of the case investigation) to threaten them. Surprisingly, 2 out of 10 respondents know about such cases. The first one is familiar with 2 cases of evidence mishandling, whereas the second has seen only 1. 3 out of 10 respondents replied with a 'no'. The rest 3 are not sure if such an abuse of evidence could happen. Two of the candidates skipped this question. The results show that at least two lawyers accept exploitation of information gathered during the digital forensic investigation, that otherwise is not reported in general.

3.6 Survey 3: privacy from general public's perspective

After successful acquisition of the exhibit, a digital forensic investigator gets full access to a suspect's data inside the acquired image. The suspect has no way to ensure that the investigators would not access her private data that is unrelated to the case under scrutiny. For example, if a person is suspected of financial fraud, then his family holiday photographs and videos, which are not related to the case should not be accessed during the investigation. Half of the investigator survey participants accepted seeing fellow investigators copying suspect's private data that is entirely unconnected to the case being investigated. Two participants from the lawyer survey agreed having knowledge of instances where the Investigative Officer threatened the suspect using data gathered during the investigation of her case.

Both these insights indicate towards serious privacy concerns for a person whose digital devices might get seized by the law enforcement agencies for some investigation. The author designed the public survey to assess the people's sensitivity about their data privacy. Moreover, the author also framed a hypothetical question that if participant's digital device gets seized by law enforcement agencies will it affect their view regarding the privacy of their data. The public survey questionnaire is divided into two subsections that are discussed below:

3.6.1 General attitude towards privacy of data and Personally Identifiable Information (PII)

The questions in this subsection target to understand how people handle their private data. What are all those files that people consider private and where do they store them. The protection of Personally Identifiable Information (PII) is another dimension of privacy in the digital world. Authors have also put some questions related to PII in the public survey.

Storage of personal information on digital devices or places

The first question in this subsection asks the participants how frequently they store their private data on digital devices that they own or use. The responses are produced in Table 3.5. The percentages stated in the table are obtained by adding up values from the responses 'sometimes', 'usually' and 'always'.

Considering the private data stored on above stated devices, losing one could be a serious privacy threat to the owner. The subsequent question asks whether respondents have lost any of their digital devices in the past five years. The responses are shown in Table 3.6. The figures show that 59.5% people have lost at least one of their digital devices in the past five years. This high number shows that either the owners are not cautious about the security of their devices or they had their devices stolen at some point in time. Valuable items like smartphones and laptops could be on target for thieves, but the loss of low-cost devices like pen drive can only be accredited to the casual behavior of the owner. People take backups of

Table 3.5: Private data stored on digital devices.

Devices	Percentage of users storing their private data
Mobile Phones	70.30
Laptops	75.10
Desktops	54.90
Portable HDD	45.40
Pen Drive	58.10

Table 3.6: Digital devices lost in the past five years.

Devices	Percentage of users who lost their respective devices
Mobile Phones	33.00
Tablets	0.70
Laptops	3.10
Portable HDD	3.30
Pen Drives	39.90
None of above	41.50

sensitive data on portable storage devices like pen drive, and the survey responses show that around 40% people have lost one in the past five years.

Common passwords for different accounts

About 32.6% of the participants use a common password or pass phrase for the security of their multiple online accounts, whereas 45% respondents use unrelated passwords for their various accounts. The rest of 22.4% participants preferred not to reveal any information in this regard. The results show the casual behavior of people towards online password security recommendations and security of data as a whole.

Storage of passwords on digital devices

The responses for this question say that 24.6% people store their passwords on either their mobiles or tablets. 25.6% people store passwords on laptop or desktop. Although two in every three, 63.9% people, do not store their passwords on their devices, the digital devices of the remaining one in every three would have their passwords stored in them. If seized for investigation one in every three devices would contain stored passwords and out of that one in every three persons would have used a common password for their accounts.

Personal files stored on digital devices

Author's aim of framing this question was to introduce the participants to a comprehensive list of private files that are stored on various digital devices they own or use. The listed files would help them appreciate what all is at risk if their devices get seized for a digital investigation. The question asks people to specify the device(s) on which they store a respective private file. Participants are required to do this for all listed private files. The responses would provide a relative ranking of the devices where people prefer to save a particular class of private files.

A total of 1474 respondents answered the question and due to space limitation in the paper, only the notable findings are enlisted in this paper. 84.27% people store their personal photographs on their laptop/desktop, and around 30 to 35% people store personal photographs on pen drives, portable hard disks, online accounts and smartphones each. The size of a digital photograph could be easily accommodated on different storage media, that makes it the most ubiquitous personal file across all digital devices and online storage services.

Other prominent files and documents stored on a variety of digital devices are stated in Table 3.7. For each type of private file specified in the question, a person's laptop or desktop stores the highest percentage of them as compared to every other device or online storage services. This finding endorses our hypothesis that an individual's laptop and desktop tend to contain a lot of private data whose privacy is at stake if it gets seized for a digital forensic investigation.

Rating personal files

The subsequent two questions were framed to obtain a relative ranking of personal files and Personally Identifiable Information (PII) respectively. The participants were asked to assign a rank to the entries on the scale of 1 to 5, where 1 is for the least important and 5 is for the most important. The motive behiend collecting these ranking was to encourage each participant to assign a relative priority to her personal file and PII data before they are introduced to the process of seizing exhibits and digital investigation. After collecting preliminary rankings in the first question, the author asks candidates in the second question if their rankings would change

Table 3.7: Personal files and documents stored on digital devices.

Type of file or document	PC	Pendrive/ Portable hard disk	Tablet	Smart-phone
Personal-photographs	81.50	33.90	6.70	30.30
Video files	69.10	23.10	3.80	20.50
Audio files	62.90	20.70	3.70	22.00
Bank-statements	43.60	7.10	1.50	4.80
Air/railway bookings	50.40	9.80	3.10	12.80
Marksheet/ admit card	67.30	15.00	3.10	7.20
CV	71.90	20.40	4.10	10.60
Medical Reports	36.30	6.30	1.80	3.10
Job offers	58.30	10.80	2.30	5.90
Passport	49.20	10.70	2.40	5.00
PAN card [1]	52.50	10.20	2.40	4.90
Aadhar card [2]	44.10	8.60	2.00	4.40
License	42.90	8.10	1.90	4.80
Voter ID [3]	46.10	8.40	1.80	4.40
Birth-certificate	45.30	8.20	1.70	3.10
Credit/Debit card details	32.60	5.20	1.20	3.70

[1]The identity card issued by Income Tax department of India.
[2]A biometric identity card issued by Government of India.
[3]The identity card issued by Election Commission of India.

assuming a hypothetical scenario where their devices get seized by the agencies.

The first question got responses from 1474 people. After adding the values of rating 5 and rating 4 for every entry, 63% of the respondents rated personal photographs as important. Other notable entries are detailed in Table 3.8.

Rating Personally Identifiable Information (PII)

This question aims to get a relative ranking among various Personally Identifiable Information (PII). 1287 people finished the survey and assigned a rank to given PII on the scale of 1 to 5, where 1 is for the least important and 5 is for the most important. 70.39% people find their full name to be important, similarly 67.66% people rated 'Father's name' as an important PII. 61.53% people consider 'Mother's maiden name' an important PII. Other notable entries are stated in Table 3.8.

Table 3.8: Ratings in descending order for personal documents/files and PII stored on digital devices.

Highest rated Personal documents/files	Participants (%)	Highest rated PII	Participants (%)
Credit/debit card	76.90	Phone number	74.60
PAN card	73.10	PAN details	72.30
Marsheet/ admit card	72.00	Email address	72.20
Voter ID	71.40	Full name	70.39
Passport	68.60	Bank details	69.90
License	68.30	Biometrics	69.10
Aadhar card	65.60	Date of birth	68.60
Personal photographs	63.00	Residential address	68.50
Job offers	62.80	Passwords	68.30
Birth Certificate	62.70	Father's name	67.66
CV	61.90	Passport details	65.60
Bank statements	61.20	Aadhar card details	64.00
		Licence details	63.60
		ATM PIN	62.20
		Mother's maiden-name	61.53

3.6.2 Awareness about digital forensics investigation

The questions in this subsection intend to understand if given a hypothetical situation where participant's digital devices get seized by the agencies, how would it change her priority ratings towards the private data.

The author expected a radical shift in a person's privacy ratings for her private data when it was in her secure custody versus when the security agencies seize the devices to investigate some case. The shift in privacy perception would be inversely proportional to the trust that people have in law enforcement agencies. The possible change in attitude would also depend on the individual's awareness about the digital forensic investigation process and the fact that most of the digital forensic tools can find hidden and deleted data.

Table 3.9: Rating when law enforcement agency acquires device.

Data Type	No Effect	May Increase	May Decrease
Personal Documents (1304-responses)	47.3%	43.8%	8.8%
Personally Identifiable Information (1304-responses)	46.7%	46.6%	6.8%

Belief in law enforcement agencies

People have firm belief that law enforcement agencies would not misuse their data, in case they seize it for an investigation. Table 3.9 show that 56.21% and 53.45% participants say it would have no effect on their previous privacy ratings. Some participants would be rather less concerned about the privacy of their data as their privacy concerns 'may decrease' after the seizing.

Awareness about digital forensics

When asked if law enforcement agencies have tools to recover deleted data, 32.21% participants are not sure if it is possible and the other 20.25% don't believe that deleted data can be recovered. Only 47.4% people know that the law enforcement agencies can recover deleted data. Even after nearly half of the people know that the deleted data is recoverable, in next question that followed 40.95% people said that they temporarily store their personal information on their office devices before deleting them.

3.7 Solution to the data privacy problem

The outcome of the surveys indicates that a suspect's data privacy is at stake during the digital forensic investigation, and there is an urgent need to incorporate data privacy measures into the process model. The solution should preserve the data privacy of suspect, but should never compromise on either the completeness of investigation or the integrity of the digital evidence. A solution that could enhance the efficiency of investigator by saving on time or effort is much desirable.

Dehghantanha and Franke [22] have defined the same as a cross-disciplinary field of research and named it as ‘Privacy-respecting digital investigation’. They also discuss the present challenges and opportunities that the field has to offer. The next subsection briefly reviews some literature that address privacy in the context of the digital forensic investigation. The subsection after that would explains author’s proposed solution approach for the problem.

3.7.1 Privacy and Digital Forensic Investigation

Aminnezhad et al. [4] states that digital forensic investigators face a dilemma whether they should protect suspect’s data privacy or achieve completeness in their investigation. The paper also states that there is a lack of awareness among professional digital forensic investigators regarding suspect’s data privacy, which could result in an unintentional abuse.

There have been attempts to protect data privacy during digital forensic investigation using cryptographic mechanisms. Law et al. [39] have proposed a way to protect the data privacy using encryption. The authors discuss encrypting a data set on an email server and indexing the case related keywords, both at the same time. The investigator gives keyword input to the server owner, who has the encryption keys, to get back the emails that contain the keyword. Hou et al. [33] proposes a mechanism to protect the privacy of data on third party service provider’s storage center using homomorphic and commutative encryption. At the same time, the mechanism also ensures that the service provider does not get to know the queries that were fired by the investigator. Hou et al. [32] illustrate a similar solution on a remote server. Shebaro et al. [67] uses Identity Based Encryption to carry out a network traffic data investigation in privacy preserving setting. Gou et al. [30] puts forward generic privacy policies for network forensic investigations. Croft et al. [21] has proposed a mechanism where data is compartmentalized into layers of sensitivity, less private data on lower layers and highly private data on higher layers. Investigator’s access to private information is controlled by initially restricting his access to the lower layers first. The investigator is required to prove his knowledge of the low-level layers, to get access to higher level information.

Van Staden [70] has proposed a Privacy Enhancing Technologies (PET) based framework to protect privacy of third parties during a digital forensic investigation. The solution requires the investigator to write focused queries while searching for potential evidence. The PET system evaluates whether the results requested by the query cause a privacy breach or not, if yes then the investigator is asked to write a more focused query. In case the investigator overrides the query results and tries to access suspect's private data, the system starts logging investigators access details in a secure place.

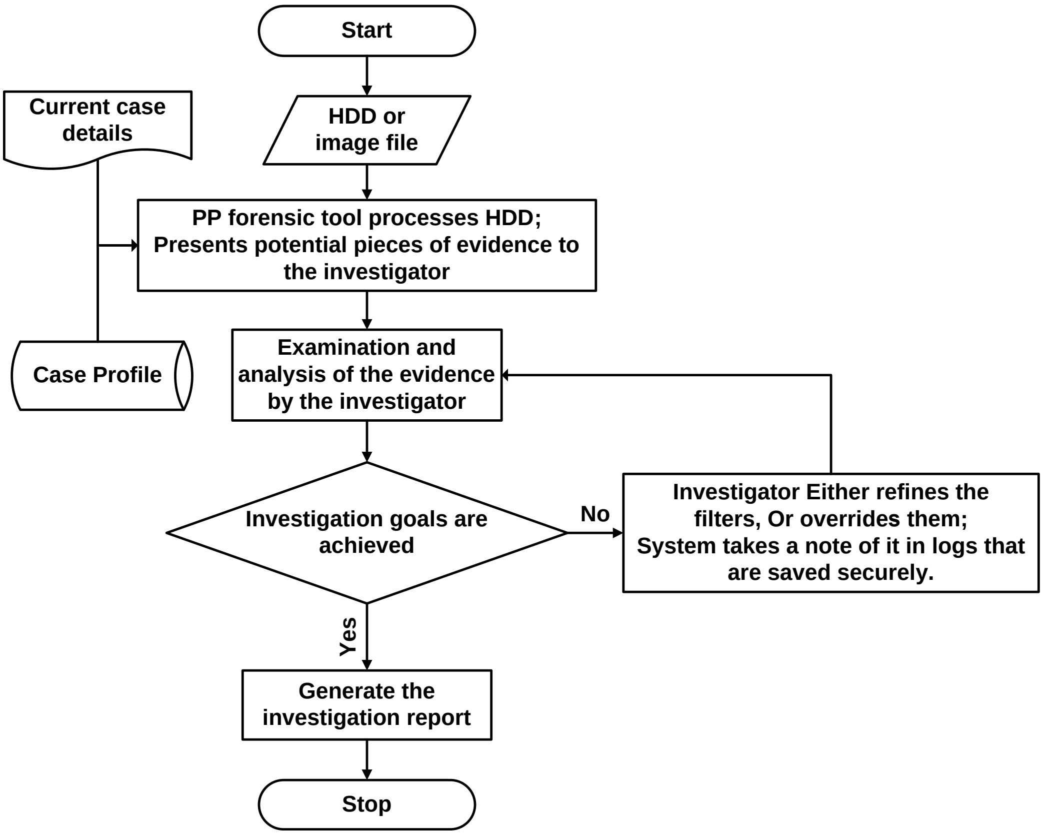

Start
Current case details
HDD or image file
PP forensic tool processes HDD; Presents potential pieces of evidence to the investigator
Case Profile
Examination and analysis of the evidence by the investigator
Investigation goals are achieved
No
Investigator Either refines the filters, Or overrides them; System takes a note of it in logs that are saved securely.
Yes
Generate the investigation report
Stop

Figure 3.1: Solution methodology flowchart.

3.7.2 Proposed solution

The author proposes a solution to protect data privacy in the digital forensic investigation process that does not interfere with the efficiency or the outcomes of the process. The solution approach (figure 3.1) brings more transparency to the process, and increases accountability of the investigators.

The solution scheme focuses on the analysis phase of digital forensic investigation where the investigator starts investigating the image or the clone corresponding to the exhibit seized. In addition to the HDD/ image-file, the solution takes two additional inputs namely the past learned knowledge of similar cases from the case profile database, and the case details of the existing case required to be investigated. The case profile database is a collection of case specific features that may be used to predict potential pieces of evidence for a particular case. The database contains feature list based on contents and metadata details of respective evidence file(s) and investigator reviews that are obtained as an outcome of learning from historical case studies. The feature list selection for such a database also needs taxonomical information about private data and files that exist on a computer system.

All the inputs are processed by the Privacy Preserving (PP) forensic tool that finds out a list of potential pieces of evidence relevant to the given case. The PP forensics tool needs to preserve the completeness of investigation. The tool could generate false positives, but should never report a false negative. Van Staden's [70] PET system asks the examiner to write focused queries to fetch potential evidence files from the image. If the tool finds investigator's query results to be violating third party privacy, then the investigator has two options. Either the investigator writes a fresh query in such a way that the query results do not intrude upon third party's privacy, or she overrides the PET filtering to carry out investigation in a conventional manner. The PET solution starts logging investigator's post overriding investigative actions at a secure storage space that is immune to all tampering attacks. In all, the PET approach adds one more layer of search without any efficiency or knowledge gain.

The proposed PP tool simplifies things for the investigator by advising potential evidence for the case in hands. If the investigator finds the results to be insufficient, then either she could mark the existing evidence and fine tune the predictions by adding more information to the PP tool's fresh run, or she could override the prediction results and continue her investigation in the old manual way. The timestamps and logs for all the activities of the investigator from the start till end, are logged in a secure manner. When the Modification Access Change Date and Time-Stamps (MACDTS) corresponding to a particular event are obtained directly from the kernel of the operating system [7] [71], the authenticity of such data is strengthened. The timestamps and logs will be very helpful when the investigator might be required to explain her actions in case any privacy breach is suspected or reported. If the investigator gets sufficient evidence from the output list to prove or disprove the case, then she may stop the further investigation and generate the case report.

The above solution will make no change to the investigative powers of the investigator. It would bring more accountability and transparency to the overall process. The investigator would get a clear idea about her responsibilities towards suspects' data privacy. There would be no compromise on performance while using this solution. However choosing correct filters/parameters so that the probability of finding all possible evidence in step 4 (before the filters are overridden) is maximized.

3.8 Limitations

It should be noted that in the public survey nearly 654 out of 1889 (34.6%) respondents did not finish the survey. Some of the participants who quit the survey in between pointed out that the survey was too exhaustive and long. Although the number of participants in the investigator survey (digital investigation experts) and the lawyer survey (cyber law experts) are limited, yet the collected responses are valuable enough due to the expertise level of these participants in their respective fields.

3.9 Summary

The paper presents the results and analysis of three surveys aimed at gathering perception on data privacy during digital forensic investigation from the three stakeholders involved, namely, the investigator, the lawyer and the general public. The analysis of results shows a lack of professional ethics among the investigators, lack of legal structure to check privacy abuse during investigations and lack of awareness in general public regarding their privacy rights.

The investigator survey results indicate towards the requirement of a better investigative model that should incorporate measures to protect data privacy without compromising on the outcome of investigation. The lawyer survey results indicate towards the requirement of privacy protection laws that fix accountability of the investigators in case of an abuse. And lastly, the general public results show that there is a lack of awareness among people regarding their personal data privacy. Although the survey has collected responses from India, the findings show valid concerns confronting the global digital forensic community.

The author has also presented a simple solution methodology to incorporate data privacy into the digital forensic investigation process. The proposed solution gives a high level approach to the given problem that does not interfere with the investigative powers of the investigator, yet able to include accountability for a possible data privacy breach by the investigator.

Next...

The author concluded that data privacy could not be incorporated into the digital forensic process as an external entity, and it should be included into the design of the digital forensic investigation model. The privacy protection should be included as a measure of transparency in the investigation process, which can also be used to fix accountability of the examiner, in case a data privacy breach is reported.

In the following chapter, the author presents the design and details of a new digital forensic framework that embodies automation to increase efficiency and still offers a reasonable level of data privacy protection. The new framework elaborates the above-stated proposed model and introduces the use of Machine Learning techniques for predicting potential pieces of evidence related to the case under investigation.

Chapter 4

DF 2.0: DESIGNING AN AUTOMATED, PRIVACY PRESERVING, AND EFFICIENT DIGITAL FORENSIC FRAMEWORK

4.1 Introduction

Digital forensic science has evolved a lot since the first Digital Forensic Research Workshop [48]. However, there are some research problems that are continuously challenging the researchers and practitioners till date.

The first and foremost challenge is the ever growing data storage capacity of digital devices [55]. The large volume of data increases the time requirements for the data acquisition and the data analysis processes [41]. Moreover, since the number of cases that involve digital evidence in some form is on the rise all over the world, the digital forensic investigators are facing a pressing need for reducing the investigation time per case [3].

The second challenge is thrown by the increasing diversity of digital devices that are becoming available in the market [31]. A digital forensic personnel has to continuously strive for finding new ways (through software as well as hardware means) to acquire and analyze such devices [35]. The software diversity deals with a huge number of file-types, ever evolving Operating Systems, the newly developed innovative applications, and other software advancements concerning contemporary digital devices. On the hardware front, diversity of sensors, chips, circuit modules and other hardware units that produce unique data streams presents a challenge for digital forensics. Although providing a solution to both of the above-stated diversity challenges takes only a one-time effort for the practitioners and researchers; however, the rate at which these parameters change keeps them on their toes.

Furthermore, people tend to use separate devices for communication, entertainment and productivity purposes. Hence the number of individuals who own and use more than one digital devices at a time is increasing [24]. Another study by Facebook in 2016 reveals that 94% teens in France and 98% teens in Germany own multiple devices [25]. The Pew Research Center published a report in 2015 stating that around 36% of US adults own all three devices, namely a smartphone, a computer, and a tablet [5]. Another survey by Pew in January 2017 has revealed that 77% of US adult population owns a smartphone, 78% owns a desktop or laptop, and 51% owns a tablet computer [50]. Although the survey presents separate figures for the three devices, one can safely assume that individuals who own multiple devices are a significant part of the US population today. The people in other regions of the world either share similar trends or would achieve the same figures in the near future. The rise in the number of devices owned per person would increase the average number of exhibits seized in a new case, thus increasing the respective investigation time and efforts.

Even after finding their ways to acquire and analyze the new digital devices, the digital forensic examiners face the third challenge from the rapidly changing technological advancements that change the rules of the game now and then [28]. The technological progress that poses a challenge to investigators is concerned with the increasing list of devices that are going digital every day, thanks to the novel software and hardware innovations. The devices in everyday use which get equipped with computational, communication and digital storage capability, commonly referred to as Internet of Things (IoT), pose new investigative challenges to the digital forensic process [47]. Any investigation involving such devices would require knowledge about how the data is produced, stored and communicated to these devices.

The fourth challenge, which is not directly connected to the functioning of the digital forensic investigation, is data privacy protection during the digital forensic investigation [4]. The digital forensic investigators always get full access to the contents of seized storage media which according to them is necessary for achieving completeness. Apart from containing potential evidence files, the seized storage media also contains owner's private data which may be sensitive at times like private/family pictures and videos, business related digital documents, medical diagnostic or treatment reports, commercial software with license information, and much more. Investigator's open access to these private files is a threat to owner's data privacy [72].

The data privacy protection is also related to need for transparency in the digital forensic investigation that ensures only case-relevant data are accessed from the seized media and remaining private files are not affected [22]. There is a pressing need for finding means to fix accountability of the investigator in case a data privacy breach happens during the investigation. The two sister agencies that work in close collaboration with digital forensic personnel, namely the Police and the regular forensic laboratories, are facing difficulties related to transparency and accountability. The case of Annie Dookhan is a good example of the same [23]. To the best of author's knowledge, there are no reported instances of professional misconduct against digital forensic investigators till date; however, it is high time that the community should adopt self-regulatory ways to improve the transparency as well as the accountability of the digital investigation process.

Apart from the challenges listed above, some researchers have predicted that moving forward the field of digital forensic would start diverging into more specialized sub-fields [29]. The sub-fields would require the investigators to get expert knowledge of the same. The digital forensic laboratories would need an investigation mechanism that could allow different experts to work together in a given case. To build a capability to handle increased number of digital forensic cases in future, the agencies would like to have prompt training programs that could prepare new and inexperienced investigators.

There is one more aspect to learning that captures the psychological, cultural and social characteristics of the people who commit crimes [59]. Researchers have been trying to capture such parameters that could help in digital forensics investigation process [58, 61].

Digital forensic frameworks to date have focused on addressing the above-stated challenges either in separation or well-defined scenarios with controlled environmental conditions. In the current chapter, the author has proposed a new digital forensic framework that incorporates forensic image preprocessing, tool-independent automation, machine learning based filtration of most relevant evidence and their privacy level evaluation to address the above-stated challenges. The framework proposes a new way in which the state of the art digital forensic research and systems could be combined in one place to realize the following.

- Increased investigative efficiency by saving in the investigation time and efforts
- Improved investigative accuracy by using multiple tools at the same time
- Better investigative planning via automation
- Improved validation
- Data privacy protection for forensically non-relevant private files
- Enhanced transparency and accountability
- Building expert knowledge for forensic investigation, education, training, and multi-agency collaborations

4.2 Proposed solution

The framework takes forensic exhibits and images (of desktops, laptops, smartphones, tablets, or other devices that store data), network logs, memory dumps, and all other sources of digital storage as input (refer to figure 4.1).

As the inputs proceed to the next phase of ***'Forensic Preprocessing'***, the investigator fills in all case related facts into a document called *Current Case Information (CCI)*. The document consists of forensically relevant data that is unique to the case under investigation, like individual keywords, timelines, and other useful information. After that, the investigator also provides the list of digital forensic tools, with their respective version numbers. All input images are processed to remove forensically irrelevant data like files listed in NSRL [66] and duplicate files [45, 65]. The forensic image formatting is also changed, without breaking the integrity of the input, to enable fast and parallel operations in successive investigation phases. In case physical devices (exhibits) are available, then the imaging for these seized devices is started simultaneously with the data removal and reformatting. The author calls the above procedure 'forensic preprocessing' as it precedes the actual processing for finding evidence files (the analysis phase). The preprocessing aims to rearrange and consolidate the data available in all of the submitted forensic images (provided in any of the popular formats) so that forensic tools could read the data concurrently. However, all preprocessing techniques and methods should ensure that the output produced by them is compatible with all digital forensic software tools. The section 4.3 discusses preprocessing in details.

The next step runs the ***'Automated Digital Forensic Processing'*** module. The module takes inputs from the CCI document, a case-specific command list, and some already known exception commands. The '*Case Profile Commands (CPC)*' database contains a list of commands that a specific digital forensic tool would require while performing a case specific job under a particular hardware deployment. These commands listed in CPC-database ensure that the planning of investigative steps is complete and consistent with respect to a particular type of case. For example, in the case of a financial fraud investigation, the CPC-database will contain commands for say Encase tool, version 7.0 running on a Windows 8.1 workstation, to perform a keyword search job (with a list of unique operations, called job-sections, refer figure 4.2) on a Linux machine's forensic-image that has an EXT4 file-system. The CPC-database contains the comprehensive collection of commands and scripts, to complete distinct tasks, which are executed by the list of forensic software tools already provided by the investigator.

The *Exception Commands (EC)* database consists of command structure similar to

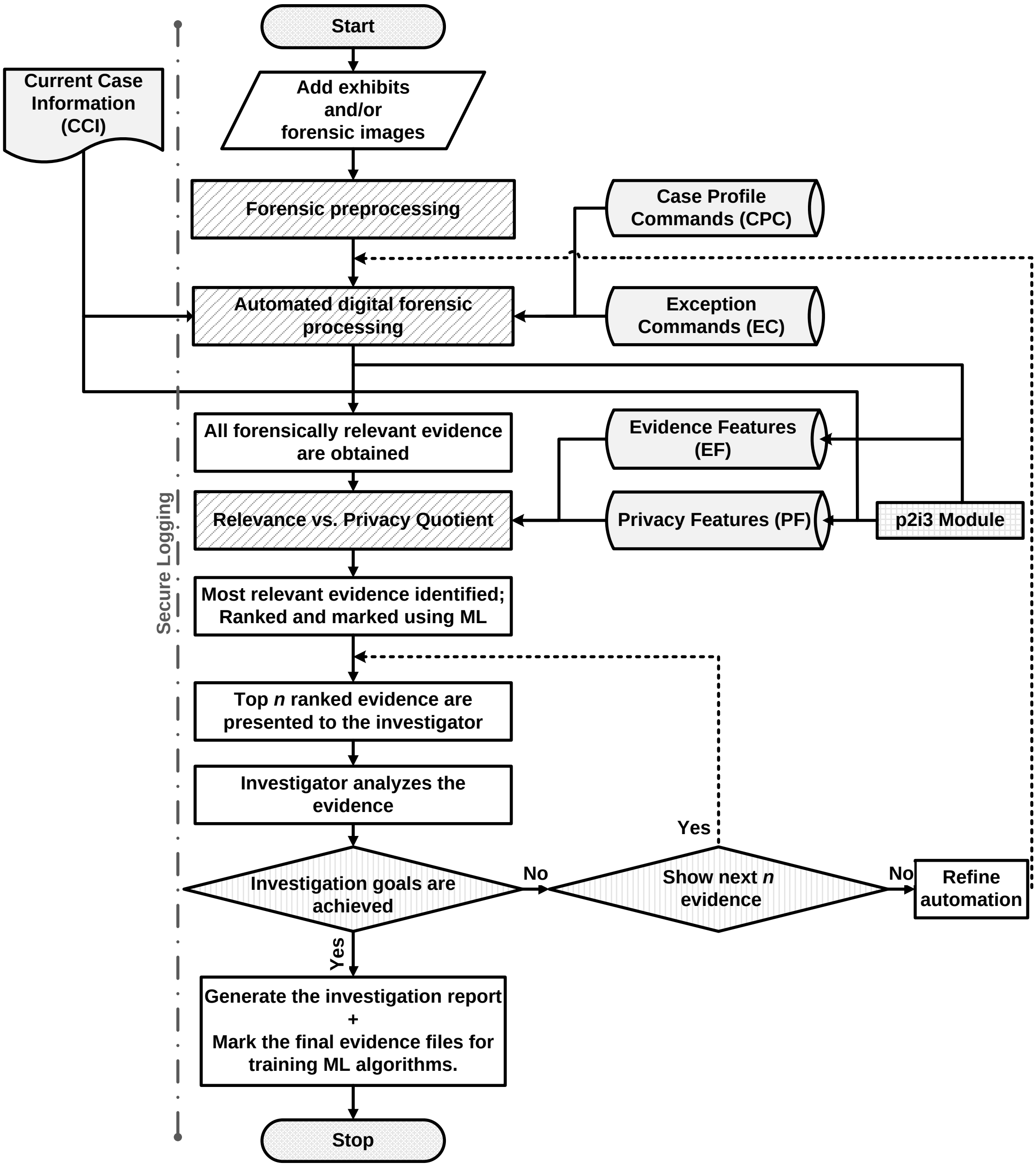

Start
Add exhibits and/or forensic images
Current Case Information (CCI)
Forensic preprocessing
Case Profile Commands (CPC)
Automated digital forensic processing
Exception Commands (EC)
All forensically relevant evidence are obtained
Evidence Features (EF)
Relevance vs. Privacy Quotient
Privacy Features (PF)
p2i3 Module
Secure Logging
Most relevant evidence identified; Ranked and marked using ML
Top *n* ranked evidence are presented to the investigator
Investigator analyzes the evidence
Yes
Investigation goals are achieved
No
Show next *n* evidence
No
Refine automation
Yes
Generate the investigation report + Mark the final evidence files for training ML algorithms.
Stop

Figure 4.1: Digital Forensic 2.0 framework flowchart.

that of the CPC-database with a distinction that these commands aim to find evidence files that could otherwise be missed during the initial run of forensic tools.

For example, all PDF attachments received on Gmail while being viewed by the receiver's browser generates one PNG image for each page of the attached document [73]. So, when the user login into their account and check their emails with PDF files as attachments, the PNG images corresponding to each page of the viewed PDF document get loaded into the browser cache. These images could be extracted from any of these three sources; the cache on hard-disk drive, the RAM dump or the Hibernation file of the system. A digital forensic investigator should fill in command (or scripts) to parse these PNG files, from the sources described above, in the EC-database.

The EC-database is a collection of all such exclusive commands which can find targeted content. In other words, the database contains expert knowledge which has been acquired over time from individual experience, careful observations, and novel research efforts. In case two forensic labs enjoy a considerable amount of trust and mutual understanding, they could share their EC-databases. The sharing will give the examiners on both sides the opportunity to upgrade their knowledge and enhance their capabilities. In case all forensic labs in a province or state agree to share their EC-databases, it could become a good collection of valuable regional (*demographic*) forensic insights.

Depending on the investigation needs and the availability of forensic tools, the automation module can work with both the open source as well as commercial digital forensic tools. The framework requires that the forensically relevant files processed by the tools have a uniformly high level of data abstraction. For example, the tools should expand all compound files (at a lower level of data abstraction) to extract the contained files (at a higher level of data abstraction) before these files could be passed on to the next level of scrutiny by the framework. Section 4.4 discusses this in more details.

The results of Automated Digital Forensic Processing are passed on to the next step (***Relevance vs. Privacy Quotient***). Here, with the help of machine learning algorithms, a relevance score for all potential evidence files (obtained from the automation module) is calculated. Similarly, the privacy quotient for these files is also calculated simultaneously. The investigator is then presented with a finite list of the top scoring relevant files. The investigator can analyze these files to decide whether these evidence files are sufficient to prove or disprove the case. If the investigator wants, she could keep on requesting the next lot of most relevant files for further examination, till the list of potential evidence gets exhausted. As soon as the investigator gets sufficient evidence from the relevance list, she may stop the investigation and generate the case report. However, if the investigator feels that the artifacts presented in relevance list are not sufficient, she is free to override the filters and start over the automation module.

The framework also incorporates a ***Secure Logging System*** (from start of the investigation till it stops) where all actions and decisions of the investigator are chronologically logged into a secure place. The safe storage for these logs could either be a hardened local server or a reliable cloud space where the investigator has no chance of tampering with them [7, 71]. Since the investigator may be required to explain her actions in case any privacy breach or some foul play is either suspected or reported. The secure logging ensures that the accountability of the investigator is fixed when such a situation arises. A brief discussion on the same is presented in section 4.6.

Automation used in the framework simplifies repeatability of the investigation process, which proves to be very helpful in validating the investigation outcomes. Especially, for the ***Technical Validation*** which aims to check whether all steps followed by the investigator fulfill the goals of the investigation. Automation together with the secure-logging will help the digital forensic community to study and optimize the investigative techniques followed by examiners. Repeatability and easy validation could improve the overall transparency of the investigative process. The framework also ensures a three-way error reduction mechanism using automation. Firstly, the automation reduces the chances of human error that may happen at any time. Secondly, the automation ensures that no step is missed from the investigative planning which remains consistent for a particular type of case. Thirdly, the automation ensures that no evidence file is missed due to limitations of a particular tool since results from different forensic tools are combined to present a comprehensive list of potential pieces of evidence. The above solution will keep the investigative powers of the investigator intact with good chances that her overall efficiency gets improved.

4.2.1 Setup

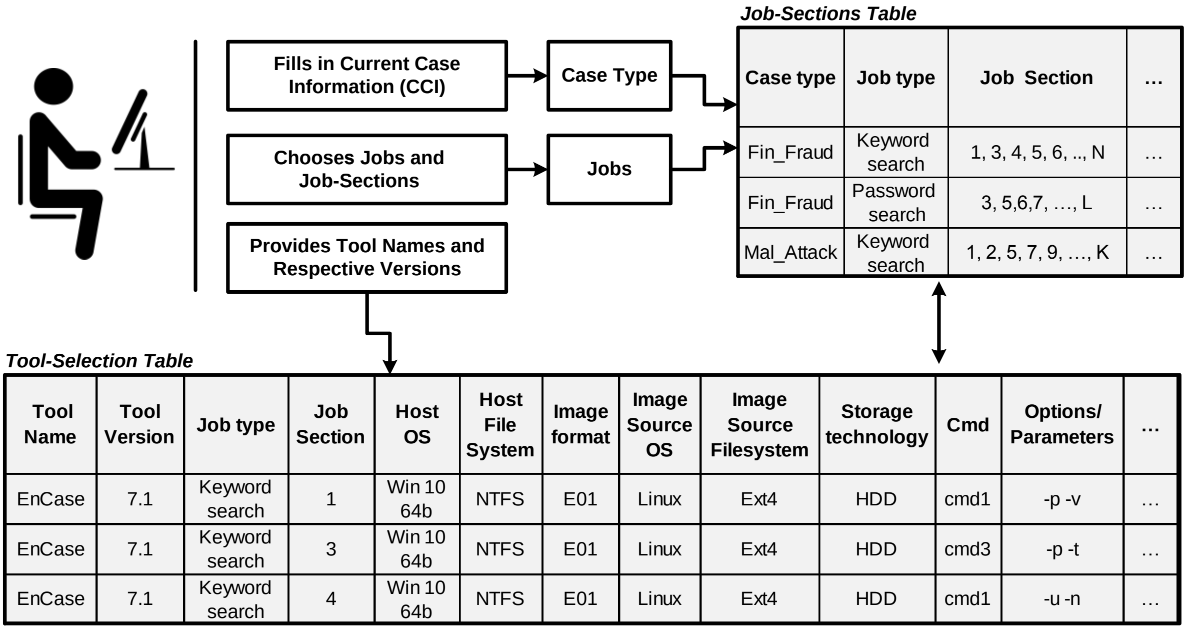
The proposed framework needs a hardware infrastructure that could provide both high-performance computational power as well as high-speed data storage and access. A robust and capable software should also support the hardware to realize both an efficient parallel processing and a powerful data management mechanism. Another requirement for the software component of the framework is its compatibility to run applications and programs from all publicly available software platforms. So, all state of the art Operating System dependent and Operating System independent digital forensic tools, which are capable of working on various digital devices, irrespective of whether they are closed source (commercial) or open source could be deployed on the proposed framework.

All the forensic tools and applications that are installed on the framework should be able to receive command-line instructions. Since most of the open source digital forensic tools take command-line inputs, they can easily be attached to the framework. Since all commercial tools are closed source, it is the responsibility of their developers to provide a command-line support for their respective tools. Although there are some tools like EnCase, which accept scripts to automate some investigative tasks, there is still a segment of commercial tools that do not support automation. The tools that do not provide any support for automation can not be used with the proposed framework.

Depending on the requirement, the proposed framework can be set up on any of the following configurations:

1. *Beowulf Cluster in a laboratory*- best suited for digital forensic laboratory environments where a suitable number of processing nodes could be selected based on the budget and investigative load [6]. A Beowulf cluster file system provides support for high-performance data access and storage. The processing speed and efficiency of a Beowulf cluster in a laboratory setting are better as compared to a distributed systems deployment or a cloud deployment of the same configuration.

2. *On the Cloud* - a private cloud with a strict access control could be a useful


Fills in Current Case Information (CCI)
Chooses Jobs and Job-Sections
Provides Tool Names and Respective Versions
Case Type
Jobs

Job-Sections Table

Case type	Job type	Job Section	…
Fin_Fraud	Keyword search	1, 3, 4, 5, 6, .., N	…
Fin_Fraud	Password search	3, 5,6,7, …, L	…
Mal_Attack	Keyword search	1, 2, 5, 7, 9, …, K	…

Tool-Selection Table

Tool Name	Tool Version	Job type	Job Section	Host OS	Host File System	Image format	Image Source OS	Image Source Filesystem	Storage technology	Cmd	Options/ Parameters	…
EnCase	7.1	Keyword search	1	Win 10 64b	NTFS	E01	Linux	Ext4	HDD	cmd1	-p -v	…
EnCase	7.1	Keyword search	3	Win 10 64b	NTFS	E01	Linux	Ext4	HDD	cmd3	-p -t	…
EnCase	7.1	Keyword search	4	Win 10 64b	NTFS	E01	Linux	Ext4	HDD	cmd1	-u -n	…

Figure 4.2: Investigator’s input to the framework.

option for an investigation agency, which has multiple departments located at the same or different geographical locations [69]. Alternatively, an agency could also rent virtual machines on a public cloud having comparatively high processing and data storage capabilities. The catch with cloud-based deployment is the dependency on limited upload and download speeds. However, if the network speeds are favorable, the cloud-hosted framework could enhance remote investigations capabilities where investigators could simultaneously work on the same case.

3. *Distributed Systems* - could also be used to deploy the framework with the processing power comparable to the above-mentioned deployment models. However, the data access speed, the parallelization in processing, and the file system capabilities would be relatively more complicated and hard to manage [57].

4.3 Preprocessing

The Forensic Preprocessing module is the first component of the proposed framework that operates on the forensic images. The author calls the module 'forensic preprocessing' as it precedes the process of finding evidence files (the analysis phase). The preprocessing aims to rearrange and consolidate the data available in all of the submitted forensic images so that forensic tools could read the data concurrently.

Before preprocessing could begin, the investigator is required to fill in all case related details into the *Current Case Information* (CCI) document. The document consists of forensically relevant information about the case under investigation, like the type of case, the name of the case, suspect's information, keywords of interest, timelines of interest, targeted file types, and other valuable information(refer figure 4.2). After filling the CCI document, the investigator also provides the list of digital forensic tools, with their respective version numbers, which are installed on her forensic system and best suit the analysis requirements of the given case. The information from the CCI document and the tools list is used by the preprocessing module to fine-tune its operations.

The primary aim of the preprocessing module is to change the data formatting of the forensic images (without breaking their integrity) so that the digital forensic tools attached to the framework could perform highly efficient parallelized operations. The secondary aim is to remove forensically irrelevant data from the forensic images which include files listed in NSRL [66] and duplicate files [45, 65].

In case physical devices (exhibits) are submitted instead of their forensic images, then the imaging for these seized devices is started simultaneously with the data reformatting and redundancy removal. All preprocessing techniques and methods should ensure that the output produced by them is compatible with the digital forensic software tools due to be used in the automation phase.

The data formatting operation should keep the integrity of the forensic images intact, and hence there should be no impact on the admissibility of the forensic evidence extracted out of the newly formatted data.

4.4 Automation

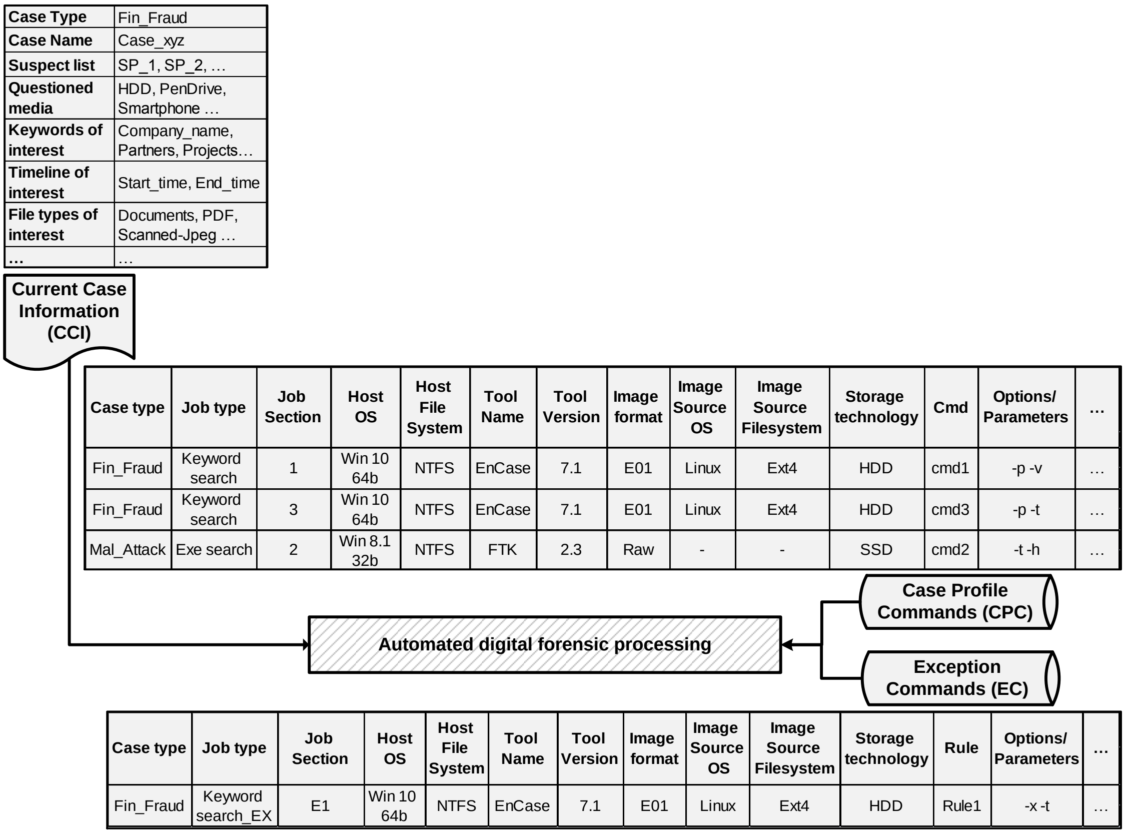
The Automated Digital Forensic Processing module aims to carry out a thorough forensic analysis of the forensic images to collect all case related potential pieces of evidence without any human intervention. The module uses the *Current Case Information* (CCI) document and queries both the *Case Profile Commands* (CPC) database as well as the *Exception Commands* (EC) database (refer figure 4.3).

The CPC-database is populated by querying two tables, namely the *Job-Sections* table and the *Tool-Selection* table (positioned at top right and bottom of figure 4.2 respectively). The Job-Sections table contains information about various jobs and subjobs (the author calls them **job-sections**) that are carried out by the digital forensic tools. The job name specifies a particular task of forensic importance which is used in a digital investigation, for example 'keyword search'. The keyword search can further be divided into small tasks, like searching keywords in all text files (let us call it job-section 1). Similarly, searching for keywords in pdf files is another sub-task (let us call it job-section 2). Likewise, a comprehensive list of well-defined subtasks for a particular job can be populated. If we consider the keyword search job with reference to a particular case (say Financial Fraud), the investigator can identify the list of job-sections that are useful for the investigation of that case.

The Job-Sections table contains this mapping for all type of known case types, respective jobs that are needed to be performed for these case types and the comprehensive list of job-sections for the same.

The Tool-Selection table contains tool version specific commands or scripts to implement job-sections from the Job-Sections table. All of the instructions are stored with respective parameters.

The CPC-database is populated with case-specific commands recognized by the tools, specified by the investigator, for completing a collection of small investigative



Case Type	Fin_Fraud
Case Name	Case_xyz
Suspect list	SP_1, SP_2, …
Questioned media	HDD, PenDrive, Smartphone …
Keywords of interest	Company_name, Partners, Projects…
Timeline of interest	Start_time, End_time
File types of interest	Documents, PDF, Scanned-Jpeg …
…	…

Case type	Job type	Job Section	Host OS	Host File System	Tool Name	Tool Version	Image format	Image Source OS	Image Source Filesystem	Storage technology	Cmd	Options/ Parameters	…
Fin_Fraud	Keyword search	1	Win 10 64b	NTFS	EnCase	7.1	E01	Linux	Ext4	HDD	cmd1	-p -v	…
Fin_Fraud	Keyword search	3	Win 10 64b	NTFS	EnCase	7.1	E01	Linux	Ext4	HDD	cmd3	-p -t	…
Mal_Attack	Exe search	2	Win 8.1 32b	NTFS	FTK	2.3	Raw	-	-	SSD	cmd2	-t -h	…

Case type	Job type	Job Section	Host OS	Host File System	Tool Name	Tool Version	Image format	Image Source OS	Image Source Filesystem	Storage technology	Rule	Options/ Parameters	…
Fin_Fraud	Keyword search_EX	E1	Win 10 64b	NTFS	EnCase	7.1	E01	Linux	Ext4	HDD	Rule1	-x -t	…

Current Case Information (CCI)
Automated digital forensic processing
Case Profile Commands (CPC)
Exception Commands (EC)

Figure 4.3: Automated digital forensic processing module.

jobs. The values obtained from the CCI document include specific terms including names of the suspects, names of the companies they are associated with, names of their partners, names of the projects they have handled, and more.

The CPC-database holds all job specific directives that may belong to more than one type of case profiles; for example, keyword search is one job which has application in a variety of cases. The keyword search job can be performed by various digital forensic software tools. However, the search technique implementation along with the keyword list(s) would differ depending on the tool specifications and the case profile respectively.

The collection of all jobs that are performed for a particular case type is in public knowledge. Moreover, how a particular job could be carried out by various digital forensic software tools could also be documented. There are tool-specific commands for performing a particular job which could take specific parameters and options based on the case type and information from the CCI document.

All of the above information is captured in the databases, as shown in Figure 4.3 that makes the automation possible. For example, if the job requirement is keyword search for a Financial Fraud case type where a Windows 10 machine with EnCase version 7.1 installed on it is available, and the forensic image is a Hard Disk Drive with Linux installation that needs to be examined, then the first database entry for keyword search could fetch the command(s) with corresponding parameters and options (if applicable). For simplicity of understanding the author has all columns of the databases in Figure 4.3; otherwise, the databases could be normalized further.

Even after processing the forensic image with a variety of digital forensic software tools, there are some crucial evidence that might escape the examiner's scrutiny. For example, with the surge in mobile phone usage people have started taking pictures of various documents that they use in their daily lives. Examples include tickets, different identity cards, business cards, bank cheques, mark-sheets and sometimes usernames and passwords for important on-line accounts. The forensic tools that search for keywords only focus on files that have textual data, and would not be able to search for images that have some written content until and unless they are instructed to do so. Experienced investigators have knowledge of such intricate details, like running OCR on suspected images along with keyword search, or filtering out the potential pictures by their metadata in case the OCR engine fails. These approaches could help the investigation by obtaining crucial evidence on the first run. The proposed framework stores these intricate details in the EC-database. The commands include implementation tricks and techniques that come from knowledge gathered by forensic experts over time as well as research breakthroughs. Structurally the database is similar to CPC-database (refer figure 4.3).

The working of the automation module (especially the structure of CPC-database) which is presented above is inspired by the work of Karabiyik et. al [36]. However, to the best of author's knowledge, the conceptualization of the Exception Commands database is a fresh contribution.

4.4.1 Design

An Expert System could be used to design the automation engine. The rules of conducting forensic analysis could be stored in the CPC-database. Different variables that need to be considered like case type, job specification, device type, respective OS and File-System versions, forensic tool's name/version, and respective commands/parameters/options could be modeled into the system.

4.4.2 Relevant vs. Non-Relevant Files: first level of data privacy preservation

The outcome of the automated digital forensic processing would give a list of files from the forensic image(s) which are potential pieces of evidence for the case under investigation.

The automation module operations segregate all files present in the forensic image(s) into two classes, namely ***Forensically Relevant Files (FRF)*** and ***Forensically Irrelevant Files (FIF)***. The FRF advances to the next stages of the investigation, whereas the FIF is made inaccessible to the investigator.

The denial of access to all files (including the private files) which are present in FIF group, is the **first level of data privacy preservation** ensured by the proposed framework.

4.5 Forensic Relevance vs. Data Privacy

The data privacy aims to protect owner's personal information from falling into hands of unauthorized people [26] [27]. Whereas, a digital forensic investigation seeks to find all potential pieces of evidence that indicate a malicious activity carried out in digital space [52].

All files that are selected/highlighted/exported at the completion of the automation module fall into the Forensically Relevant Files (FRF) group. The number of files in the FRF is still large enough for the investigators to examine individually. Moreover, a considerable number of owner's private files that do not qualify as concrete evidence are also included in the FRF collection. Hence, finding actual evidence files from the FRF group is undoubtedly a massive manual effort, which further involves a significant risk of data privacy violations for the private files that do not have much of evidential value.

The proposed framework uses machine learning to determine the degree of relevance (details in subsection 4.5.1) as well as the level of privacy (details in subsection 4.5.2) for all files present in the FRF group. The investigator is presented with the top most relevant files (say, a bunch of top 20 or top 50) for examination, with their respective level of privacy also marked on them.

The next set of most relevant files is not presented to the investigator until she examines the first bunch and feels that further investigation is needed. Only after the investigator raises an explicit request to the system, the next bunch (succeeding 20 or 50) of files is presented for her scrutiny. The process of request and grant continues until the investigator finds all actual evidence needed to resolve the case or the list of FRF gets exhausted. In a rather unusual situation when the examiner feels that the automation module should be rerun, the framework provides a provision of doing so too.

The above-stated mechanism, for presenting most relevant files in a bunch until the investigator finds concrete evidence to prove or disprove the case, also prevents privacy breach to an extent. The process could also be understood as the **second level of data privacy preservation** which is ensured by the proposed framework. Although the data privacy protection in this filtration process is not absolute, however, the data privacy of a large number of files belonging to FRF is significantly preserved.

4.5.1 Degree of relevance

The proposed framework classifies files based on their degree of relevance to the current case under investigation. The classification process needs to process data available in the Evidence Features (EF) database (Figure 4.4). The EF-database takes information about each file that is selected into FRF, and some case specific information from the Current Case Information (CCI) document.

Feature selection

The aim is to classify each file in the FRF into a potentially-conclusive or a potentially-indecisive piece of evidence. The information stored in the EF-database corresponding to each file, belonging to the FRF for a particular case under investigation, acts as a feature-set for a machine learning implementation. The features can come from:

1. The file's metadata: includes information like - File-Type, Time-Stamps, File-Size, File-Address, File Containing Folder Name, File Containing Folder Depth, Access Control Permissions, and Owner(s) of the File

2. Source image and the automation module: includes information like - Forensic Tool that selected the file, More than one Tool selected the file (Y/N), Job-Type, Job-Section, Level of Data Abstraction, Did the file get extracted from a compound file (Y/N), Source Image Format, Source Image File-System, Source Image Operating-System, Source Image Storage Technology

3. Use of the Exception Commands: includes information like - Is a result of Exception Command (Y/N), Number of Exception Commands used, Exception Command IDs

4. The associated Current Case Information: includes information like - Case-ID, Case-type, Has Keywords of Interest (Y/N), Has Name from Suspect List (Y/N), Is File Type of Interest (Y/N), Does Fall into Timeline of Interest (Y/N)

It is worthy to note that, the list of above-stated features is not exhaustive and may contain more features in each category. Also, the order of features mentioned above does not reflect their respective significance.

Data collection

The data collection happens when a case is investigated using the framework. Two options that may be used by the investigating agencies while doing the data collection are discussed below:

1. Data collection for a particular type of case: It includes collecting data while investigating cases of the same kind. For example, If an investigative agency analyzes only Financial Frauds cases, then all features collected in the Evidence Features database will help in forming a machine learning prediction model most suited for financial fraud cases. Creating a model for a particular kind of case is considerably easy because each case shares a high degree of commonality in their respective feature sets.

2. Data collection for all type of cases: It includes collecting data while investigating cases of all kinds. The features collected in the Evidence Features database will form a machine learning prediction model that could find potentially conclusive evidence for any given case. Creating a generic model that can make predictions for any case at hand is a difficult task as compared to the previous option because the feature sets will have many variations.

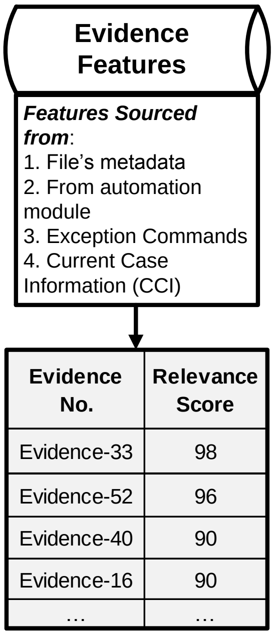
Machine learning approach for relevance

As already stated before, the machine learning solution aims to classify each file in the FRF into either a ***Potentially Conclusive (PC)*** or a ***Potentially Indecisive (PI)*** evidence. Hence, to put it formally -

1. The machine learning approach addresses a two-class classification problem (a *supervised learning technique*). The reason for choosing a supervised learning approach is to learn from the experience of the investigators who have already solved similar cases. The framework needs access to the case related artifacts like the case information document, the forensic image associated with that case, information about the tools that were used to solve the case, and the list of actual evidence files that concluded the investigation.

The first three artifacts (*mentioned in the previous paragraph*) could be used by the framework to collect feature information about all the FRF files, whereas the last object would act as the ground truth for training. All actual evidence that the investigator marks at the completion of each case investigation help populating the last feature column that is helpful in training.

After training on some examples of solved cases of the same type, the machine learning solution could start predicting for a new case. However, for a generalized solution, the training set should contain a considerable number of examples of each type of case that have been solved by the investigative agency before the solution could start predicting.



Evidence No.	Relevance Score
Evidence-33	98
Evidence-52	96
Evidence-40	90
Evidence-16	90
…	…

Figure 4.4: Degree of relevance for forensically relevant files.

2. The supervised learning approach could be implemented using an ensemble learning method like Decision tree or Random Forest that give considerably good results when the training data set is less, and the feature set is relatively strong.

> The author thinks the above-stated learning methods are suitable for the classification task (PC vs. PI) when developing a prediction model for the same type of cases with a relatively small training dataset. However, if an investigation agency that has a collection of a substantial number of cases of the same type say hundred or more cases of financial fraud, then they could try other algorithms like Support Vector Machine (SVM) and k-Nearest Neighbors (kNN).
>
> When a generic solution needs to be created, an ample number of cases of each type that the investigation agency works on is required. However, if multiple agencies agree to share their EF-databases and list of conclusive evidence for respective cases, the aim of making a generic prediction solution could be achieved.

The machine learning approach finds PC files and calculates a relevance score for each of them. The files are then arranged from highest relevance score to the lowest. The framework ensures that only a bunch of most relevant files are presented to the investigator and rest of the files are masked from her. The investigator asks for the next bunch of files if required. The process continues till the investigator finds all conclusive pieces of evidence or the list of FRF gets exhausted. The machine learning solution's efficiency increases with the number of solved cases getting incorporated into the training set.

As explained in the *sub-subsection 4.5.1* the potentially conclusive evidence are can be presented for the investigator's scrutiny using the following algorithm:

Algorithm 1 Evidence examination

```
for $k = 1$ to $p$ do
    Pick $S_k$
    for $l = 1$ to $m$ do
        if $F_l$ is PC-Evidence then
            Bookmark $F_l$
            break
        else
            continue
        end if
    end for
end for
```

4.5.2 Privacy quotient

The framework also identifies whether a file is private or it contains any Personally Identifiable Information (PII) about the suspect. The aim is to correlate the data privacy information for each file with their respective evidence rating (from the previous subsection). The privacy information of each file will not restrict the investigative capabilities of the forensic examiner in any way. However, the privacy quotient of the individual file would enable both the suspect and the legal authorities to assess the scale of data privacy violation, if it happens during the investigation process.

A specific module named Private and PII Identification (**p2i3**) runs on all files belonging to FRF (refer figure 4.1). The author has marked the p2i3-module as a separate entity in the flow diagram; however, the module could be a part of the automation engine if some of the forensic tools support the required functionality. For example, the tool EnCase (version 7 and up) has the provision of finding files that contain personal information as well as artifacts containing Personally Identifiable Information.

All files in the FRF group are examined to determine whether they are private to the suspect or contain any of her PII.

Feature selection

The information stored in the Privacy Features (PF) database acts a feature-set for machine learning implementation to find each file's privacy quotient. The features are described below:

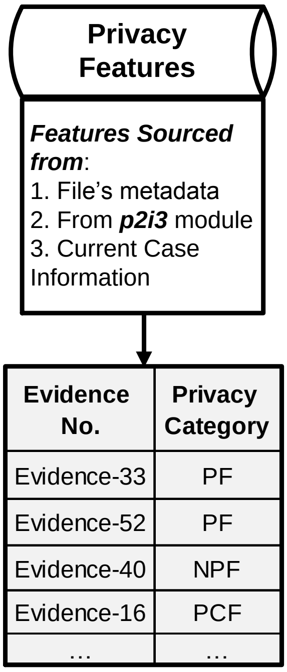
1. Features from file's metadata (same as in the EF-database): It captures information like - File-Type, Time-Stamps, File-Size, File-Address, File containing folder name, File containing folder depth, Access-Control permissions, Owner(s) of the file
2. Features from the source image and the **p2i3** module: It captures information like - Source image format, Source image File-System, Source image Operating-System, Source image storage technology, Is the file a private file (Y/N), Type of

the private information identified, More than one type of private information present (Y/N), Does the file contain any PII (Y/N), Type of PII identified, More than one PII present (Y/N)

3. Features from the CCI document: it captures information like - Case-ID, Case-type, Has keywords of interest (Y/N), Has name(s) from the suspects list (Y/N), Is the File-Type of interest (Y/N), Does the file fall into Timeline of Interest (Y/N)

It may be noted that, the list of above-stated features is not exhaustive and may contain more features in each category. Also, the order of the features does not reflect their respective significance.

The data collection part of the privacy rating solution is same as that of the evidence rating solution (refer sub-subsection 4.5.1).



Evidence No.	Privacy Category
Evidence-33	PF
Evidence-52	PF
Evidence-40	NPF
Evidence-16	PCF
…	…

Figure 4.5: Privacy quotient for forensically relevant files.

Machine learning approach for privacy quotient

The aim of the machine learning implementation in the privacy solution is to categorize files from the FRF group into three groups; namely, the Private Files (**PF**), PII Containing Files (**PCF**), and Non-Private Files (**NPF**). Hence, to put it formally -

1. The machine learning approach addresses a clustering problem. An unsupervised machine learning approach is used to categorize the files into one of three clusters (*PF, PCF, and NPF*) as described above.

2. The unsupervised learning approach can use a k-means algorithm to segregate the files into these three clusters. However, there are good chances that the third cluster NPF could get more than 35% of sample population (files from FRF), making the k-means cluster analysis unfruitful. In such a situation the solution needs one extra level of processing.

   The k-means algorithm should be started with a higher value than n (the number of required clusters, currently n=3); preferably 3 to 4 times the value of n.

   The result of the previous step would give 9 to 12 clusters, each of which would comply with the condition of having the sample population between 5 - 35%. A secondary level of clustering on top of these results (using Hierarchical Clustering) can be used to club them into the final three clusters namely, PF, PCF, and NPF.

4.6 Secure logging system

The logging process ensures that all operations from the starting state in the proposed framework (refer the flowchart in figure 4.1) till the state when the investigation stops are recorded. The logging also ensures that all actions of the examiner starting from the time when she begins the analysis process till all conclusive evidence get identified are listed. All system operations and investigator actions need logging because of two reasons – firstly, to resolve conflicting situations like allegations of data privacy violations – secondly, for studying investigation styles of examiners for learning and training purposes.

The logging system could fulfill both of the above-stated requirements only when the logs are complete as well as tamper-proof. The first requirement of completeness, which is relatively easy to achieve, refers to logging all activities of the system and the investigator.

However, the second requirement of ensuring that the logs become tamper-proof is a difficult problem. The first possible solution could capture the activity logs with the help of a dedicated application running on the forensic system. This solution assumes that the examiner is cooperative and honest enough not to interfere with the logging application. After the investigation process is complete, the logging application should transfer the logs to an external storage place which is safe from tampering. Any tampering attempt during its operation would cause the application to stop prematurely, invalidating the captured logs.

The second possible solution should try to capture examiner's activities at the operating system level (with a system level application or module) and save the logs in a safe location. The safe storage for these logs could either be a hardened local server or a reliable cloud space where the investigator has no chance of tampering with them [7].

Since the investigator may be required to explain her actions in case any privacy breach or some foul play is either doubted or reported, the secure logging fixes the accountability of the investigator for her actions, in case such a situation arises.

4.7 Summary

The author has proposed a new digital forensic framework that brings efficiency in digital forensic processing with the help of automation while preserving data privacy for the suspect. The framework ensures that the automation supports a range of digital forensic software tools and produces effective outcomes by incorporating the current case information, case profile data, and the knowledge of experienced digital forensic investigators. The investigator is presented with the most relevant evidence that are sorted with the help of Machine Learning algorithms. The framework balances the investigative requirements of the case with the data privacy protection of suspect's forensically irrelevant private files.

The framework ensures the data privacy protection of the non-evidential private files at two levels. Firstly, by denying access to all files that are marked as Forensically Irrelevant at the end of the automatic analysis module. The secondary level of data privacy protection is provided to those private files which get a low forensic relevance score from Machine Learning classification algorithm. These files are not presented to the investigator for inspection, hence keeping their privacy intact. Although the data privacy protection at the secondary level is not absolute, however, the data privacy of a large number of files belonging to FRF is significantly preserved.

The framework also ensures that the efficiency of investigation is enhanced, without compromising on the outcomes of the investigation or affecting the investigative powers of the examiner. However, since the system is securely logging all actions of the investigator, she experiences a greater sense of accountability for avoiding unwanted data privacy violations. The automation and secure logging encourage a better validation check, hence bringing a higher level of transparency into the investigation process.

Next...

The next chapter presents a prototype implementation of the Machine Learning algorithms for predicting the potential evidence to the investigator and determining the privacy quotient of those files.

Chapter 5

PREDICTING EVIDENTIAL-VALUE AND EVALUATING PRIVACY-QUOTIENT FOR DF 2.0: THE MACHINE LEARNING APPROACH

The author presents a prototype (*proof of concept*) implementation of Machine Learning (**ML**) techniques for predicting the evidential value of a file (section 5.2), as well as assessing the privacy quotient of each file (section 5.3).

Before going into the implementation details, the following section illustrates the mathematical formulation to explain how the relevance score prediction could be expressed as a ML two class classification problem.

5.1 Mathematical formulation of relevance score in ML

Let the number of input cases be ***n***, and the number of features corresponding to an individual file be ***x*** (*from the EF-database, Section 4.5.1*).

$$\mathbf{C} = \{C_1, C_2, C_3, ..., C_n\}$$

Where, **C** represents the case vector. The case instance C_i can be represented as a collection of its respective Forensically Relevant Files group (FRF).

$$C_i = \{F_{i1}, F_{i5}, F_{i7}, \ldots, F_{ij}, \ldots\}$$
$$where,\ i \in (1\ to\ n)$$

And, F_{ij} is the j^{th} file in C_i's FRF. Every file in the above set can have a maximum of x features, and the feature vector for F_{ij} can be represented as:

$$\mathbf{f}_{F_{ij}} = \{f_{ij}^{1}, f_{ij}^{2}, f_{ij}^{3}, ..., f_{ij}^{x}\}$$

So, the case C_i together with its FRF and respective feature vectors can be represented in matrix form as:

$$C_i = \begin{bmatrix} F_{i1} \\ F_{i2} \\ F_{i3} \\ . \\ . \\ F_{ij} \\ . \end{bmatrix} = \begin{bmatrix} f_{i1}^{1} & f_{i1}^{2} & f_{i1}^{3} & f_{i1}^{4} & f_{i1}^{5} & . & f_{i1}^{x} \\ f_{i2}^{1} & f_{i2}^{2} & f_{i2}^{3} & f_{i2}^{4} & f_{i2}^{5} & . & f_{i2}^{x} \\ f_{i3}^{1} & f_{i3}^{2} & f_{i3}^{3} & f_{i3}^{4} & f_{i3}^{5} & . & f_{i3}^{x} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ f_{ij}^{1} & f_{ij}^{2} & f_{ij}^{3} & f_{ij}^{4} & f_{ij}^{5} & . & f_{ij}^{x} \\ . & . & . & . & . & . & . \end{bmatrix}$$

The input cases ground truth evidence can be represented as

$$E = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ . \\ E_i \\ . \\ E_n \end{bmatrix}$$

And, E_i accounts for the evidence vector corresponding to the i^{th} case which was declared solved after finding files having conclusive evidence. For example, the evidence vector will have a collection of files like

$$E_i = \{F_{i1}, F_{i3}, F_{i5}, \ldots\}$$

$$where,\ Files\ in\ E_i \subset Files\ in\ C_i$$

Here, the feature vector corresponding to the evidence E_i would consist of the union of all prominent features of files mentioned above.

$$\mathbf{f}_{E_i} = f_{F_{i1}} \cup f_{F_{i3}} \cup f_{F_{i5}} \cup \ldots$$

Let us assume that the features which get selected are following

$$\mathbf{f}_{E_i} = \{f^1_{i1}, f^5_{i1}, f^9_{i3}, f^{15}_{i5}, f^{19}_{i3}, f^{21}_{i5}, \ldots, f^x_{i1}\}$$

Since we have ***x*** input features, hence the weight vector **W** can be represented as

$$W = \begin{bmatrix} W_1 \\ W_2 \\ W_3 \\ . \\ . \\ . \\ W_x \end{bmatrix}$$

and,

$$W = funtion_1(FeaturesMatrix, EvidenceVector)$$

The Relevance Score (*RS*) for each file present in FRF can be computed as

$$RS = function_2(WeightVector, FeaturesMatrix)$$

The computation of *RS* is followed by sorting of the Potentially Conclusive (PC) files from the highest relevance score to lower. The files get clustered into various sets, say **p** number of sets, and each set has **m** number of files which can be represented as

$$S = \{S_1, S_2, S_3, \ldots, S_k, \ldots, S_p\}, and$$

$$S_k = \{F_1, F_2, F_3, \ldots, F_l, \ldots F_m\}$$

5.2 Prediction of evidential value: Classification

In the absence of a real-world digital forensic case, the author decided to choose the 'Hacking-Case' for the prototype implementation. The Hacking-Case files are available on NIST's *Computer Forensics Reference Data Sets* (CFReDS) project website [18].

5.2.1 The setup

The author downloaded the EnCase images (*two .E01 files*) from the Hacking Case page on CFReDS website. The rest of the steps are enumerated below:

1. The author collected the metadata information about all the files contained in the forensic image of the given case. The author used the EnPack *'flat-file-export-(v4-0-0).enpack'* [37] in EnCase V7 to export around sixty six columns of metadata information corresponding to each file. The table 5.1 provides selected fields produced by the above-mentioned EnPack. The author collects all metadata values in a CSV-file and name it as All-File-Dataset (**AFD**).

2. The author asked five digital forensic investigators working in a private digital forensic laboratory to find answers to 20 investigative questions (*refer Appendix B*) out of the 31, which are mentioned on the website. The author reduced the number of questions for simplifying the analysis process for the investigators. The author asked each investigator to mark a set of files as potential evidence. The author collected all five sets of marked evidence, where one investigator's set may have a slight difference in the number of entries as compared to entries in the marked sets of her colleagues. The above-stated

scenario is ideal to collect all potential evidence files for the given case because the union of all marked sets from multiple investigators would provide a comprehensive list of answers for each investigative question.

3. The author asked the investigators to align their tools (EnCase) with the time-settings of the image before they start looking for answers. The same time settings would ensure that marked files collected from all the investigators have consistent time values.

4. The investigators were asked to note down the total time they spent on the case during the investigation.

5. After the investigation process got over, the author asked the investigators to export the metadata information of their respective marked evidence files (using the *flat-file-export-(v4-0-0).enpack*) into respective Potential Evidence Dataset (**PED**).

6. The author first collated all values generated by the five investigators in one place and removed the multiple entries. In other words, all PEDs were merged into one CSV file, and duplicate rows were removed. The author named this file as All-Potential-Evidence-Dataset (**APED**). The rows in APED are unique and present the union of all the files that were marked by the investigators as potential pieces of evidence.

7. The author added a new column to APED named '*IsEvidence*'; which contained a binary value '1' for all the rows signaling that all entries in the table were potential evidence files.

8. At the same time, the author also added 'IsEvidence' column to the AFD, and made all entries '0'; asserting that all entries in the AFD were non-evidence files.

9. Finally, the author merged all entries of the APED into the AFD, and removed the duplicate rows where the 'IsEvidence' value was '0'. The final CSV became the dataset which was used in ML implementation.

It is worthy to note that some files, marked by the investigators which were registry values, while exported to their respective PED's did not have any of their timestamps (*like Accessed, Acquired, Created, Modified, or Deleted*) except Written. The author populated the missing timestamps of these registry entries using the time details of their parent files in the final dataset.

5.2.2 The dataset

As already mentioned in the previous subsection, the author has used *flat-file-export-(v4-0-0).enpack* [37] to export all files present in the Hacking-Case EnCase images. There were a total of 12,190 files present in the case. After exporting the metadata of all these files using the EnPack, the author found that only 11, 937 entries were populated in the output file (AFD). The exporting script ignored 98 files which have a physical size of zero bytes, and 120 archive/composite files were also skipped by the script during the process.

Table 5.1: The reduced set of columns present in the dataset.

Type	**Columns**
Time	Accessed, Acquired, Created, Deleted, Modified, Written
String	Category, Description, EvidenceFile, Extension, Extraction Status, ItemType, Name, ShortName, PrimaryDevice, Protected, Signature, SignatureResult, SignatureTag
Numeric	ExtentCount, FileID, InitializedSize, LogicalSize, PhysicalLocation, PhysicalSector, PhysicalSize, UniqueOffset
Addresses	FullPath, Matching File Path, OriginalPath, SymLink
Alpha-Numeric	GUID, MD5Hash, SHA1Hash, StartingExtent
Binary	HasAttributeList, HasPermList, IsCompressed_B, IsDeleted_B, IsDisk_B, IsDuplicate_B, IsEncrypted_B, IsFolder_B, IsHardLinked_B, IsHidden_B, IsIndexed_B, IsInternal_B, IsMountedVolume_B, IsOverwritten_B, IsPicture_B, IsSparse_B, IsStream_B, IsVolume_B, WasProcessed_B, **IsEvidence**

The author intentionally removed one row from the database, which had the address ‘Dell Latitude CPi\C.’ Encase adds all images in a case under an imaginary ‘C’ folder, which acts as the root folder for that case. Since the ‘C’ folder entry does not have any evidential value or actual existence in the real case, the author decided to remove the same. Hence, the total count of entries in the AFD database decreases to 11, 972.

The author then carried out the EnCase processing on case image and exported the metadata again. The processing of the image recovered metadata entries corresponding to 120 archive/composite files that were missed earlier.

After combining the metadata entries (PEDs) obtained from the five digital forensic investigators, the author got a total of 259 metadata entries for all marked potential evidence files in the APED.

It may be noted that the forensic investigators also marked registry entries as evidence files, whose metadata information are not present in the initial forensic image. The EnCase expands the registry files (like SAM) and enables the investigator to mark the entries within. There are a total of 23 registry entries that are included in the APED.

Finally, the author merged the entries from the APED into the AFD, and obtained the final dataset, with a total of 12,115 entries.

The number of columns exported by the EnPack is 66; however, the author removed some of the columns that hold information specific to EnCase or the EnPack. For example, columns like 'Codepage', 'Complexity', 'Entropy', 'Tags', and 'Recepiant' are EnCase specific columns which are not so tightly related to the actual file. Similarly, columns like 'Output filename' is an example of the EnPack specific column which is not related to the file.

Moreover, there are some other columns that hold redundant information; like the columns 'Full Path', 'Item Path', and 'True Path' have the same content. The author removed such types of columns as well, and reduced the set of columns to 55 (refer table 5.1 for details). The 56th column 'IsEvidence' is populated by the values received from the digital forensic examiners. A'1' in the column means that the file is potential evidence in the given case, and a '0' means it is not.

5.2.3 Experiments and results

The author conducted experiments on the dataset using various baseline algorithms, promising Machine Learning (ML) algorithms suited for two class classification, results of which are presented in the next section.

Digital forensic investigation process aims to capture all potential pieces of evidence, and could not afford to lose any possible evidence that may slip through the investigator's scrutiny as a benign file. The similar scenario happens in the ML results, where the False Negatives (also called Error type 2 in ML) are the files that are actually potential evidence files but have been wrongly predicted as innocuous. Considering the harm that False Negatives values can have on the outcome of the investigation, the author used the technique of 'Bagging' to reduce their values. Results of the same are presented in the next section.

Experimental protocol

The author divided the dataset into training and testing in the proportion of 80% and 20% respectively. Hence the training dataset contains 9,692 records, and the testing dataset includes 2,423 records.

5.2.4 Baseline performance

The author has used some popular ML algorithms which are known to be good performers on the two-class classification problems. The author has used seven algorithms, namely - Support Vector Machine (SVM), Two-class Logistic Regression, Deep SVM, Decision Forest, Decision Jungle, Boosted Decision Tree, and Neural Networks. There is no specific reason for the selection of these seven algorithms, and other algorithms can also be used on the dataset to accomplish the required classification job.

The confusion matrix of each of these algorithms is presented in table 5.2. The positive labeled entries (marked evidence files) in the dataset are significantly less than the negative labeled entries (where 'IsEvidence' value is '0'), so the accuracy values

Table 5.2: The baseline performance of various ML algorithms on the dataset.

Algorithm	Confusion Matrix		Confusion Matrix Format		Accuracy	EER
Support Vector Machine	28	28	TP	FN	98.37	0.1071
	12	2379	FP	TN		
Two-Class Logistic Regression	36	20	TP	FN	98.81	0.0954
	9	2382	FP	TN		
Two-Class Locally-Deep SVM	7	49	TP	FN	97.67	0.2857
	8	2383	FP	TN		
Two-Class Decision Forest	0	56	TP	FN	97.75	0.1045
	0	2392	FP	TN		
Two-Class Decision Jungle	0	56	TP	FN	97.75	0.2592
	0	2392	FP	TN		
Two-Class Neural Network	44	12	TP	FN	97.96	0.0954
	38	2353	FP	TN		
Two-Class Boosted D-Tree	44	12	TP	FN	99.31	0.0276
	5	2386	FP	TN		

of the algorithms do not reflect each algorithm's actual performance. Hence, the author has also incorporated the *Equal Error Rate* (**EER**) of the respective algorithms in the results (table 5.2). The lower the EER, the better the performance. The ROC

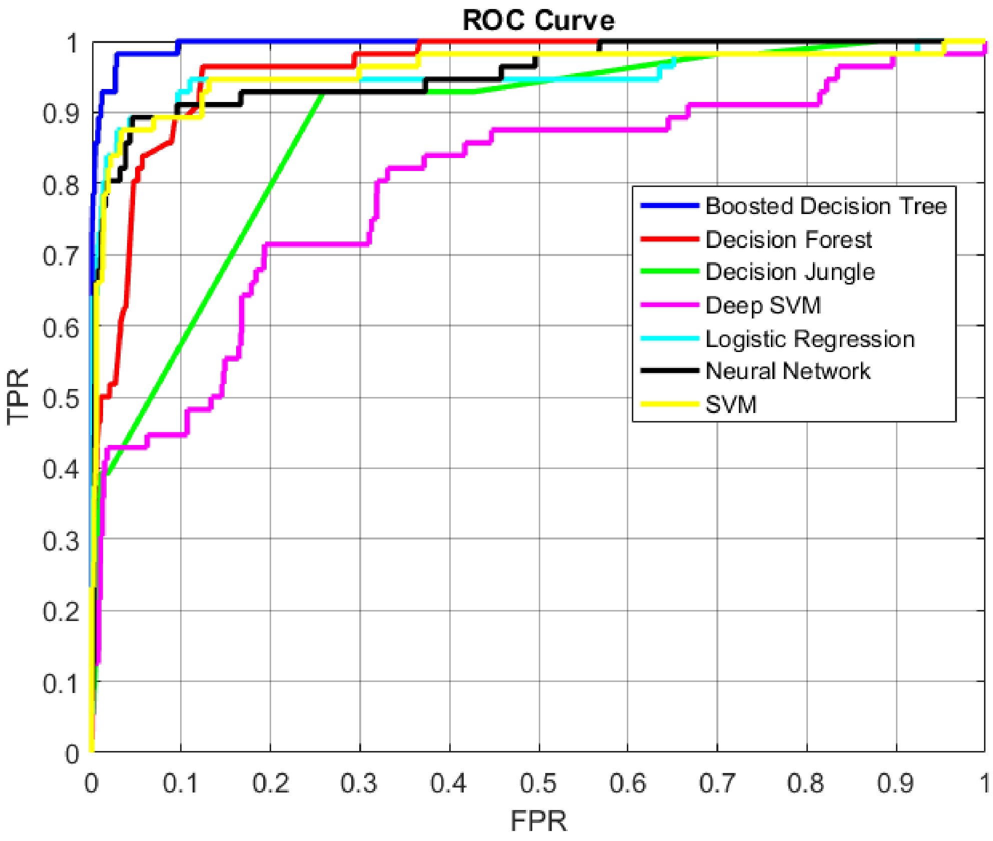


Figure 5.1: The ROC for baseline algorithms.

curves of all these algorithms are plotted in figure5.1 for easy comparison.

The baseline algorithms results show that all algorithms are not performing good when it comes to tackling the type 2 errors; the False Negatives (**FN**). The high values of the FN are not right from the digital forensic perspective too, as they allow actual evidence files to slip through the investigator's scrutiny as innocuous files. However, the False Positives (**FP**), called type 1 errors in ML terms, on the other hand, could also be problematic for digital investigator as they mark innocent files as potential evidence files. However, since all the files predicted by the proposed ML solution are presented to the investigator for final decision making, all the FP would be easily identified at that time.

The ML prediction could be meaningful in digital forensics scenario if the FN are reduced to a minimum. The author has applied 'Bagging' technique to achieve the same goal; which is explained in the subsection 5.2.6.

5.2.5 Relevance score

As mentioned in the mathematical formulation section 5.1, the framework would present the potential evidence files, sorted in order of their relevance score, to the investigator. The framework uses the 'scored probabilities values' as the relevance score for a particular file (*the table 5.3 shows some examples of the same*).

In the table 5.3, the second column '*IsEvidence*' is the labeled column which belongs to the input dataset (AFD). The third column contains the scored values predicted by the trained ML model (*here, the SVM model*). The fourth column holds the respective probability scores with which the ML model has predicted the classification.

It can be observed from the table that the entry 5, is a False Positive **(FP)**; a non-evidential file marked as potential evidence. Whereas, the entries 7 and 8 are False Negatives **(FN)**; actual evidence files marked as benign ones. The baseline implementation results show large numbers of FN, which are not suitable for the digital forensic investigation, the author resorted to 'Bagging' technique to reduce the

Table 5.3: The relevance scores of files.

S. No	IsEvidence	Scored Labels	Scored Probabilities
1	1	1	0.9996124506
2	0	0	0.0007406375
3	0	0	0.0027789336
4	1	1	0.9765605330
5	**0**	**1**	0.9028829932
6	0	0	0.0036880332
7	**1**	**0**	0.1220335960
8	**1**	**0**	0.0965592340

same.

5.2.6 Bagging

The ML technique ‘Bagging’ solves a given problem by creating multiple weak ML models that take almost equal portions of positive and negative label samples for training.

Once ready, all these ML models give their predictions for a given test sample. The

Table 5.4: The performance of ‘Bagging’ on the dataset.

No. of Classifiers	Confusion Matrix		Confusion Matrix Format		Accuracy	EER
15	49	3	TP	FN	91.5	0.0599
	203	2168	FP	TN		
30	50	2	TP	FN	91.79	0.0627
	197	2174	FP	TN		
45	50	2	TP	FN	90.3	0.0585
	233	2138	FP	TN		
60	47	5	TP	FN	91.62	0.0606
	198	2173	FP	TN		
75	51	1	TP	FN	86.83	0.0611
	318	2053	FP	TN		

final decision on that sample is taken through a majority voting over all of these predicted values.

The author took a 40 to 60 ratio of positive to negative labeled samples to train the groups of two-class classifiers using the Neural Networks ML algorithm. The author tried with five different group sizes, namely 15, 30, 45, 60 and 75 – and tested these groups of classifiers to predict for the current case. The results of these 'Bags' of classifiers are provided in the table 5.4. The ROC curves corresponding to bag-level results are presented in the figure 5.2.

It can be observed from the confusion matrix of these groups that the False Negatives (**FN**) have been reduced to low values; for example, the FN for the 75-Classifiers group is just 1. Although the accuracy value for the classifier groups varies, the au-

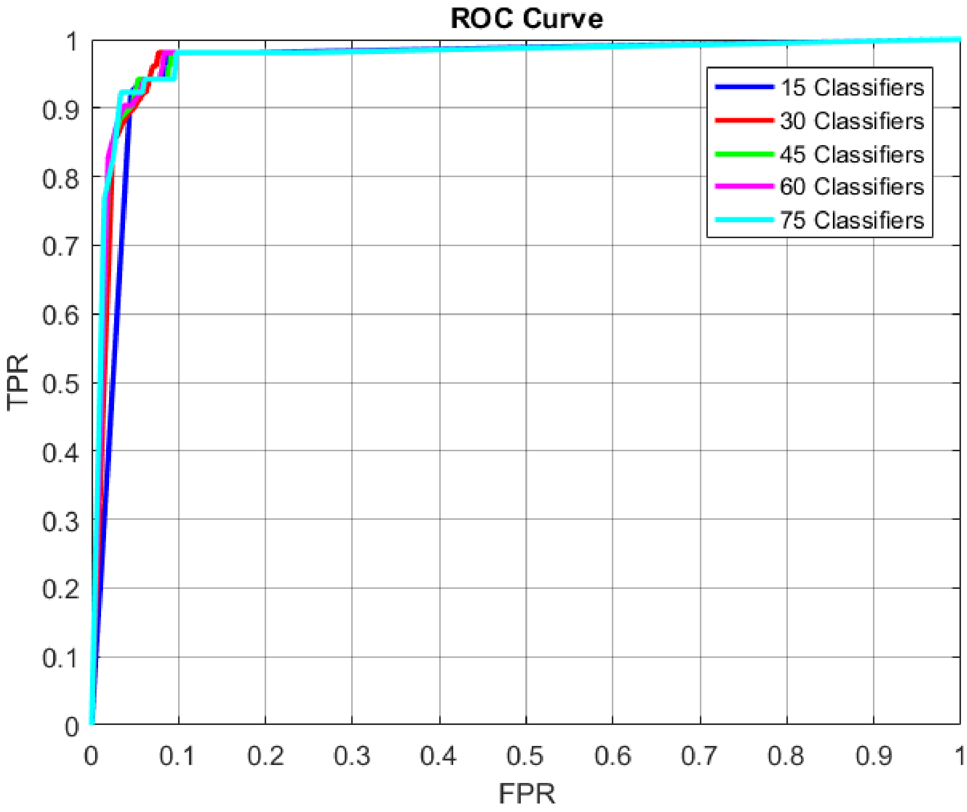


Figure 5.2: The ROC for the bagging approach.

thor observed that the 45-Classifiers group gives the best performance in terms of a low FN (2), a low EER (0.0585), and reasonably high Accuracy (90.3).

5.3 Determining the privacy quotient: Clustering

After creating a prototype ML model for predicting the evidential relevance of files, the author also implemented another ML model that could cluster files based on their privacy quotient.

The clustering implementation aims to segregate all files present in the input digital forensic image into different classes. These classes can then be labeled as either Private Files (**PF**), PII Containing Files (**PCF**), or Non-Private Files (**NPF**).

5.3.1 Dataset

The author has used the same digital forensic image as discussed in the section 5.2; which is the 'Hacking-Case', available on the CFReDS website [18].

The author has used the same dataset for the privacy ML prototype implementation. Since the 'Hacking Case' is a generated case which has been developed by CFReDS for training purposes, it does not have much private information that could be clustered out.

Hence, for the sake of simplicity and prototype implementation, the author has assumed all media files (pictures and multimedia category) as PF, all documents files as PCF, and rest of the files as NPF.

5.3.2 Dataset processing

The ML clustering algorithms use higher levels of numerical calculations in the background, before they could assign a clustering label to given entries; hence they prefer more numeric valued columns in the input datasets.

The current dataset (*introduced in the previous sub-section 5.2.2*) has a plentiful of string-valued columns like 'Category', 'Extention', 'Full-Path', 'SignatureTags', and others. So, in order to get fruitful clustering results, the author carried out data-manipulation and transformation for several non-numeric columns. For example, the author changed the string-valued binary columns (*containing 'TRUE' or 'FALSE' values*) into binary-valued columns (*'1' for 'TRUE', and '0' for 'FALSE'*). The columns in the binary category of table 5.1 with names like 'IsDisk_B', 'IsDuplicate_B', and 'IsPicture_B' are examples of the same.

For feeding data to clustering algorithms, the author dropped some columns from the input dataset which were not helping with the clustering process. The insights about which columns could be dropped, and which data-manipulation and transformation techniques should be used came from extensive experimentation.

Table 5.5: The data translation of the 'Category' column.

Numeric - Code	**Categories**
1	Library
2	Windows
3	Executable
4	**Picture, Multimedia, Multimedia-Video**
5	**Document, Document-Presentation, Document-Spreadsheet**
6	None
7	Folder
8	Archive
9	Script, Unknown, Email, Database, Communication, Plug In, Internet, Code, Font, Application

5.3.3 Experiments and results

The author used K-Means and Hierarchal clustering algorithms for grouping the PF, PCF, and NPF files. The author used a data transformation on the 'Category' column of the dataset, which maps numerical values (1 to 9) to the string values of the column. The above-stated mapping is provided in the table 5.5.

For the k-means clustering to work, every potential cluster should have between 5% to 35% of the sample population respectively. As stated earlier, the current experiment aims to group samples in three clusters, namely PF, PCF, and NCF. Since the number of samples in NCF are more than 35% of total samples, the author segregated the NCF into seven sub-groups (7 numeric codes in total). Also, since all media files are in PF, the author assigned one numeric code (*code-4*) to them. All the documents files are in PCF are assigned one numeric code (*code-5*). Therefore, the author chose nine numeric codes for data translation as mentioned in table 5.5. The segregation keeps the labeled samples in check and clustering algorithm is able to perform better.

Table 5.6: The K-Means clustering results.

Purity of cluster	Dominating class
0.677364865	1
0.999285204	**4**
0.553819444	1
0.487112046	2
0.461617195	2
0.321135991	6
0.993489583	7
0.682042834	1
0.707509881	1

All the clustering experiments aim to get a maximum purity value for code-4 and code-5 groups. The 'purity' of a particular cluster with respect to an input class refers to the probability with which the samples of that class map into that cluster after clustering. The higher the purity value, the better is the clustering result for that class.

The results of the k-means algorithm are shown in table 5.6. It can be noticed that the purity of code-4 (PF files) in cluster 2 is '0.9992' (*very close to 1*). However, the clustering results for code-5 (PCF) are not so good, as there is not a single output cluster where code-5 dominates the results.

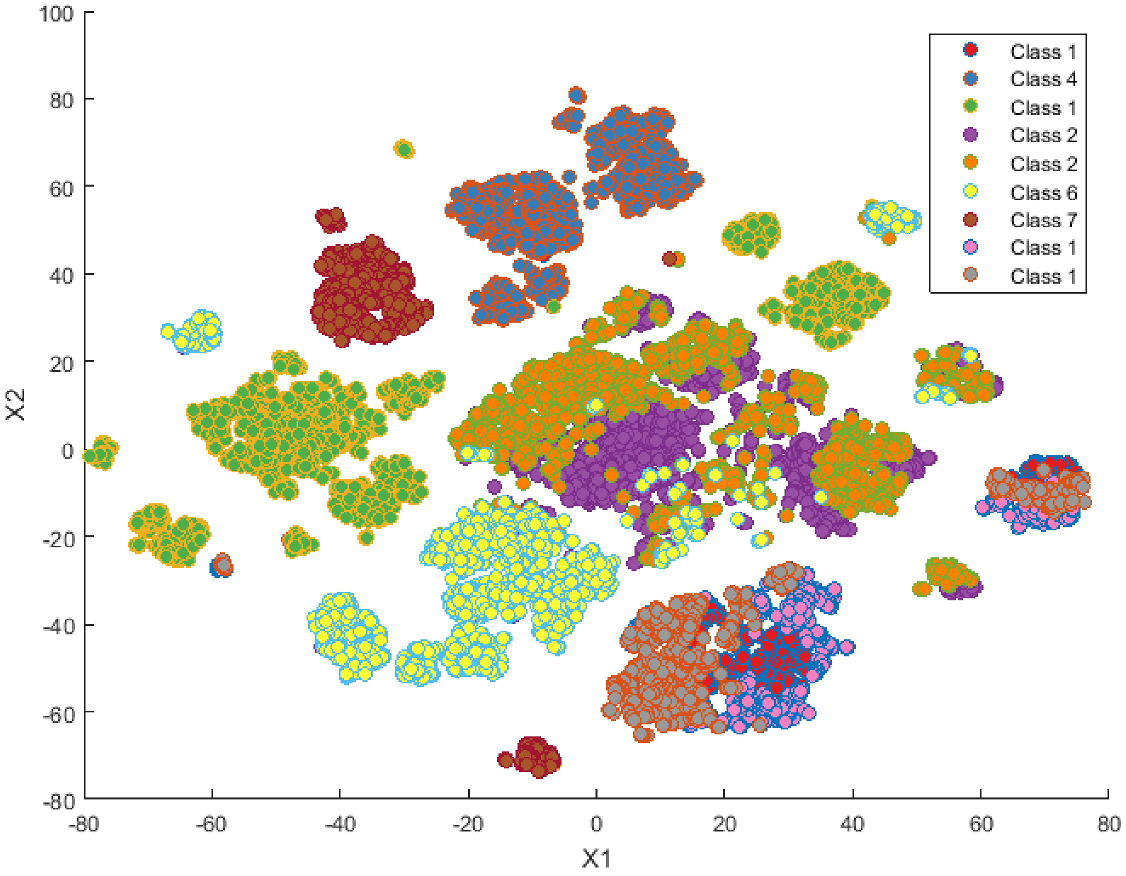


Figure 5.3: The t-SNE plot for K-Means clustering results.

The same results can be visualized in a better way using the **t-SNE** plot; which is shown in the figure 5.3. Here, the second cluster which is dominated by code-4 (class-4) can be seen as a dark patch in the top center of the plot.

However, when the authors applied hierarchical clustering on the same dataset, the results were not so encouraging. The hierarchical clustering was not able to clearly distinguish either code-4 (PF) or code-5 (PCF) in any of the nine output clusters. The results for hierarchical clustering are stated in table 5.7.

5.4 Summary

The current chapter of the thesis exhibits a prototype Machine Learning (ML) implementation that predicts the evidential relevance of a given file that is present in the forensic image of a case under investigation. The algorithm predicts whether a given file is potential evidence or not. The prediction task has been modeled as a supervised learning problem (two-class classification) where the ML algorithm aims first to get training on a labeled dataset, followed by making predictions on

Table 5.7: The Hierarchical clustering results.

Purity of cluster	**Dominating class**
1	9
1	9
1	7
0.923076923	9
1	9
1	9
1	9
1	9
0.254199420	1

the records of the testing dataset.

Firstly, the performance of seven baseline ML algorithms was tested on the CFReDS's 'Hacking Case' dataset [18]. In spite of giving a reasonable accuracy the results from these seven algorithms show a high rate of False Negatives, which is not acceptable in the digital forensic investigative scenario.

So, the author used the 'Bagging' technique, that takes the predictive decision by taking a majority voting over the predictions of a bunch of weak machine learning models that are trained on small portions of equal parts of positive and negative labeled samples from the dataset. The use of bagging significantly reduced the number of False Negatives, making the ML predictions more usable for a digital forensic investigator.

The implementation of ML techniques for assessing the privacy quotient showed encouraging results. The k-means algorithm implementation produced an exclusive output cluster that was dominated by the PF class. However, the results for the PCF class were not so promising.

Chapter 6

CONCLUSION

In the current chapter, the author summarizes the work presented in the thesis and provides some future directions.

6.1 Summary

The author started with the idea of the current research work with a question that if the investigator gets complete control over the seized storage media of a suspect in a case, then how can one ensure that only the files that have a connection with the case are accessed, and the private files of the suspect that have nothing to do with the case are not touched. There are digital forensic researchers who have answered the above-stated question by proposing solutions that focus on the data privacy protection of the suspect. However, these solutions either work in isolation or are constrained to a particular context. There was a need for a solution that gets incorporated in the current digital forensic investigation process and doesn't interfere in the investigative powers of the examiner. The author started working on the above-stated research gap.

The first and foremost problem that the author faced was the absence of ground-truth about the need for such a solution. So, the author started with a study that is aimed at gathering perceptions of data privacy during the digital forensic investigation from the three stakeholders who are involved namely, the investigator, the lawyer, and the general public (*discussed in chapter 3*). The investigator survey results indicate towards the requirement of a better investigative model that should incorporate measures to protect data privacy without compromising the outcome of the investigation. The lawyer survey results indicate towards the requirement of privacy protection laws that fix the accountability of the investigators in case of abuse. Lastly, the general public survey results show that there is a lack of awareness among people regarding their data privacy, which is the reason that there is no demand coming up from the general public to change the current situation. Although the surveys have collected responses from India, the findings show some valid concerns confronting the global digital forensic community too.

The study results also re-confirmed the author's assumption that a data privacy protection solution cannot be an external entity in the digital forensic process, and hence, it should be incorporated into the design of the digital forensic investigation model itself. Moreover, the privacy protection would act as a measure of transparency in the investigation process, which can be further used to fix accountability of the examiner, in case a data privacy breach is reported. However, if the privacy protection has to be incorporated into the current digital forensic investigation model, the efficiency of the overall process gets affected. So, there was a need for re-imagining the digital forensic investigation model which can retain or possibly increase the current efficiency of the investigation, incorporate data privacy protection, and while accomplishing the previous two ensure that the investigative powers of the investigator remain intact.

The author includes the above-stated requirements in the form of a new digital forensic framework that brings efficiency in digital forensic processing with the help of automation while preserving data privacy for the suspect to a considerable extent (*presented in chapter 4*). The framework ensures that the automation supports a range of digital forensic software tools, and, produces effective outcomes by incorporating the current case information, case profile data, and the knowledge of experienced digital forensic investigators. The investigator gets the most relevant pieces of evidence that are sorted with the help of Machine Learning (ML) algorithms (*implementation details are discussed in chapter 5*). The framework keeps the investigative requirements of the case unaffected while protecting the data privacy of suspect's forensically irrelevant private files to a considerable degree.

The framework also ensures that the efficiency of investigation is enhanced, without compromising on the outcomes of the investigation or affecting the investigative powers of the examiner. Moreover, since the system is securely logging all actions of the investigator, she experiences a greater sense of accountability and would be encouraged to avoid unwanted data privacy violations. The automation and secure logging encourage a better validation check, hence bringing a higher level of transparency into the investigation process.

6.2 Future work

The author would like to extend the Machine Learning (ML) solution in real life digital forensic cases. The solution will require access to the forensic images of real-life solved cases, which are available in various digital forensic laboratories in India, or any other country in the world. The dataset which can be deduced from these already solved real-life cases will be used for the training of ML models. Since the above-stated kind of sharing of data may be difficult for the agencies, who possess the forensic images of these real-life solved cases, for various organizational as well as legal reasons. So, the author would like to make an independent standalone system for the digital forensic personnel, who are working in a laboratory environment, for extracting the datasets from the real-life cases. These datasets can then be used to build respective ML models for predicting new cases of the same type.

The datasets will hold only the metadata, which is non-confidential information, of the files contained in the respective cases. The non-confidential nature of extracted datasets will encourage the forensic personnel to share their respective datasets and the ML trained models with the research community as well as their colleagues in other digital forensic laboratories.

In the long term, the author would like to combine the datasets from the same type of cases from different laboratories in one geographical region and train an ML model using the same. Similarly, new ML models can be trained on the datasets gathered from a group of laboratories from a state, or a country, or countries around the world. The experiments will be useful in finding how the prediction pattern varies; and how these learnings could be used in making a universal prediction system.

Another exciting extension could be applying the prediction models trained on data in one language to predict cases containing different languages. The natural language processing engines used in the digital forensic tools have to be language specific, however, the file metadata that is collected for making the dataset would be the same. So, the ML model that will be trained on the above-stated dataset would work.

Since the *General Data Protection Regulation* (**GDPR**) has come into force from May 2018, the author would also like to incorporate the privacy compliance measures into the DF 2.0.

Appendix A

THE SURVEY QUESTIONNAIRES

A.1 Questionnaire one: digital forensic expert's views on investigation

Dear participant, we are conducting this study to understand how a digital forensic expert proceeds into the investigation of a given case. A secondary aim is to understand the impact of a digital forensic investigation on privacy of data. In particular, how privacy of data owned by victim/ accused is handled during the investigation process.
All the responses collected in this study would be COMPLETELY ANONYMOUS and will be used for RESEARCH PURPOSE ONLY. The answers will NOT be shared with any third party.

* Marked are compulsory questions.

*1. When a case comes for investigation in which a digital media (PC/ laptop, mobile phone, tablet or PDA) is involved, do you follow the procedures specified in chain of custody (COC) form?

Yes, always.

Yes, most of the times.

Yes, only sometimes.

No.

Don't know about chain of custody form.

2. If your answer to the above question is 'a', 'b' or 'c', then

I follow the chain of custody (COC) procedures only if the case is going to court of law.

I follow the chain of custody (COC) procedures only in an internal investigation of corporate cases.

I follow the chain of custody (COC) procedures if I think the case is important.

I follow the chain of custody (COC) procedures in all the cases, both while going to court of law and being done for corporations.

Other (please specify)

*3. After you are done with IMAGING (a bit by bit copy of storage media), what all inputs EXCEPT 'keywords' do you choose to find potential pieces of evidence? (You may choose multiple options)

Time stamps of files.

Size of file.

Type of file (file extension).

Folder depth (in which the file is stored)

Other (please specify)

*4. Have you come across any tool that asks for a case profile (takes information/ story about the case from the case file as input) and the list of questions needed to be answered during the investigation; and decides the keywords to be searched by its own?

YES

NO

*5. At what time do you stop your investigation? (Please choose one option)

After you have gathered a minimum amount of evidence needed to prove or disprove the given case.

After you have gathered all the evidence present in the storage media that are related to the given case.

After you have gathered all possible evidence present on the storage media that include those related to the case as well as those that are not related to the case, but can be used to prosecute the owner of the media in a new case.

Other (please specify)

*6. Have you ever experienced a situation where, by chance you as an investigator get hold of some proofs/ potential evidence for other activities not related to the case, which can be used in forming another separate case against the accused/ victim?

Yes, always.

Yes, most of the times.

Yes, only sometimes.

No, never.

*7. During the investigation of a case while examining the storage media image, we assume there would be some instances when you come across owner's private files (like personal photographs, videos, business plans, or other intellectual property). Then...

You would view all such files, but copy only those which are related to the case under investigation.

You would view all such files, copy files which are related to the case under investigation as well as some other files which are not related to the case but are illegal in nature.

You would view all such files and copy all such files because these files have more probability of becoming evidence files in all cases (including the case in hand and all other possible cases).

Other (please specify)

*8. Have you seen any forensic investigator, who while investigating a given case, copies files like wallpapers, songs, movies, software games or commercial software from the case image (bit by bit copy of storage media) under investigation? (Please choose one option)

Yes, in my own company/ lab.

Yes, but not in my company/ lab.

No, but there should not be any problem in doing that.

No, and I think no one should do like this.

*9. The digital forensic labs (especially the government ones) collect all pieces of evidence for a given case as per questions asked by the Investigative Officer (IO) and on the basis of investigator's experience. All of these evidence found during the investigation process, are written to an external storage media like CD or DVD and given back to the IO. Have you seen any case during your career as a digital forensic expert, where the IO was accused of misuse of this information (the evidence CD/ DVD) to threaten the victim or the accused?

Yes

No

Don't know about such process

If yes, can you remember number of such cases?

10. In case, you do a corporate case investigation (which has least chance of going to court of law), depending on your experience what percentage of cases required you to go to their premises to carry out the investigation process?

Less than 10 %

10 to 30 %

30 to 50 %

50 to 70 %

Above 70 %

11. In case, you do investigation of cases that go to court of law, depending on your experience what percentage of cases required you to go to accused /victim's premises to carry out the investigation process?

Less than 10 %

10 to 30 %

30 to 50 %

50 to 70 %

Above 70 %

*12. According to your experience till date, does the average time taken to solve a digital forensic case depends on:
(you can choose multiple choices)

Priority of the case.

Size of storage media.

Technology of storage media.

Availability of new up to date Digital forensic tools.

Other (please specify)

*13. The organization you are working with can be categorized as:

Privately owned.

A Government body.

Other (please specify)

*14. Please specify your experience in handling cases involving digital devices (Like desktop, laptop, mobile phones, tablets and PDA etc.) in number of years:

Less than 2 years

Between 2 to 5 years

Between 5 to 10 years

More than 10 years

*15. Please specify how many Digital Forensic cases (involving digital devices) have you solved/ handled during your career till date:

Less than 10 cases

Between 10 to 30 cases

Between 30 to 50 cases

Between 50 to 70 cases

Between 70 to 100 cases

More than 100 cases

*16. Do you have a graduate or postgraduate degree in computer science or allied areas?

Yes

No

Other (please specify)

*17. In your career as a digital forensic investigator, how much of training programs have you attended to boost your digital forensic skills?

3 months

6 months

1 year

More than an year

NONE

Other (please specify)

18. Would you like to share some problem(s), you face (or have faced) during digital forensic investigation which could not be solved because there is no mechanism or tool capable of doing so?

A.2 Questionnaire two: legal view on digital investigation

Dear participant, we are conducting this study to understand the legal status of privacy of data, especially during a digital forensic investigation. A secondary aim is to know the affect of digital forensic investigation on privacy of data. In particular, how privacy of data owned by victim/ accused is catered during the investigation and court trial process.
All the responses collected in this study would be COMPLETELY ANONYMOUS and would be used for RESEARCH PURPOSE ONLY. The data will NOT be shared with any third party.

* Marked are compulsory questions.

*1. In case of a Cyber Crime and Computer Fraud case, at what time do you think your preparation of a case is complete?
(Please choose one option)

After you have gathered a minimum amount of evidence needed to prove or disprove the given case.

After you have gathered all the evidence present in the storage media of digital devices that are related to the given case.

After you have gathered all possible pieces of evidence present on the storage media of digital devices including evidence related to the case as well as those that are not related to the case, but can be used to prosecute the owner of the media in a new case.

Other (please specify)

*2. According to your experience as a lawyer, how many number of evidence (on average) are present in a minimal set of evidence which are sufficient to prove or disprove a Cyber Crime and Computer Fraud case in the court of law?
(Please choose one option)

1 or 2.

3 to 5.

6 to 10.

More than 10.

*3. Have you ever experienced a situation where the investigator, while investigating a particular case, gets hold of some proofs/ potential evidence for other activities not related to the case which can be used to file a new separate case in favor or against the accused or victim?
(Please choose one option)

Yes, always.

Yes, most of the times.

Yes, only sometimes.

No, never.

Other (please specify)

4. If your answer to the above question is ‘b’ or ‘c’, please fill the approximate percentage of such cases out of the all the cases that you have solved or handled till date?
(Please choose one option)

0 to 10%.

10 to 30%.

30 to 50%.

More than 50%.

Other (please specify)

*5. How many cases, in your career as a lawyer till date, have you handled in which a PC/ laptop’s Hard Disk Drive or Smartphone/ tablet/ PDA etc. were seized and the accused or victim have applied for ‘right of privacy’ referring to either the freedom of speech and expression under Article 19(1) (a) or right to life and personal liberty under Article 21 of the Constitution of India, or both?
(Please choose one option)

10 cases.

10 – 30 cases.

30 - 50 cases.

50 and above.

Please specify the absolute number, if you remember...

*6. Have you ever seen a case(s) in your career where the accused or victim, who’s PC/ laptop’s Hard Disk Drive or Smartphone/ tablet/ PDA etc. were seized, accused the investigative agencies for privacy breach under section 72A of the (Indian) Information Technology Act, 2000. [i.e. he/she accused the agencies for accessing and disclosing their private information, which is irrelevant to the case being investigated. Example, the access and/ or disclosure of personal/ family photographs and videos, when the person is being investigated for a financial fraud.]
(Please choose one option)

No.

Yes.

If yes, how many such cases have you seen so far...

*7. Have you ever seen a case(s) in your career where the accused or victim, who’s PC/ laptop’s Hard Disk Drive or Smartphone/ tablet/ PDA etc. were seized, accused the investigative agencies for privacy breach under section 43A of the (Indian) Information Technology Act, 2000. [i.e. he/she accused the agencies for improper or negligent handling of their sensitive personal data or information during investigation of the case]
(Please choose one option)

No.

Yes.

If yes, how many such cases have you seen so far...

*8. The digital forensic labs (especially the government ones) collect all pieces of evidence for a given case as per questions asked by the Investigative Officer (IO) and on the basis of investigator's experience. All of these evidence found during the investigation process, are written to an external storage media like CD or DVD and given back to the IO. Have you seen any case during your career as a digital forensic lawyer, where the IO was accused of misuse of this information (the evidence CD/ DVD) to threaten the victim or the accused?

Yes

No

Don't know about this process

If yes, can you remember number of such cases?

*9. According to your experience, how many cases including those you have solved and others you know about, the accused or victims asked/ requested the court to preserve their private data or files on their digital and mobile devices?
(Please choose one option)

up to 10.

10 to 20.

30 to 50.

60 to 80.

More than 90.

NONE, till date.

Other (please specify)

*10. You work:

As an independent consultant.

With a privately owned company.

With a government body.

Other (please specify)

*11. Your experience as a Cyber Crime and Computer Fraud lawyer (in number of years):

0-2.

3-5.

6-8.

8-10.

10 and above.

*12. How many Cyber Crime and Computer Fraud cases have you solved/ handled till date
(Please choose one option):

Up to 10 cases.

Up to 25 cases.

Up to 50 cases.

Up to 75 cases.

More than 75 cases.

13. Do you have any suggestion for how we can improve our survey, please specify:

A.3 Questionnaire three: understanding privacy of data

Page 1, *Title: General Attitude Towards Privacy*
This survey aims to gather user views on privacy aspects under certain assumptions. It will take 15 minutes on an average to complete this survey and the information collected in this survey will be used for educational and research purposes only.

Every participant, after successful completion of this survey stands a chance to win exciting prizes (portable hard disks, pen drives and more).

* Marked are compulsory questions.

*1. How often do you use the following devices/ places to store your personal or sensitive information? (Personal information is any information related to you that you would not like to become public)
CHOICE COLUMN - Never - Rarely - Sometimes - Usually - Always - DON'T HAVE ANY
OPTIONS -

Mobile

Tablet

Laptop

Online Social Network (OSN) like Facebook etc.

Desktop

Portable Hard Disk

Pen drive

*2. Have you lost any of the following in past 5 years?

Mobile

Tablet

Laptop

Portable Hard disk

Pen drive

NONE OF THE ABOVE

Other (please specify)

*3. Do you use a common password/ pass-phrase for different accounts/ websites over the Internet?

Yes

No

Prefer Not to say

Other (please specify)

*4. Do you store passwords for accounts/ websites on any of the following devices? (You may mark more than one)

Mobile

Tablet

Laptop

Desktop

OTHERS

NOT AT ALL

Other (please specify)

*5. Would you consider sharing your PASSWORDS for the following accounts with anyone (friends/ family/ colleagues)?
CHOICE COLUMN - Never - Rarely - Sometimes - Usually - Always - N/A
OPTIONS -

Email Account (eg. Gmail)

Social Account (eg. Facebook)

Professional Account (eg. LinkedIn)

Net Banking account

Ticket Booking Sites (eg. IRCTC)

Other account(s)

Others, Please specify below

Page 2, *Title: Privacy of Stored Data*

*6. Do you have the following documents stored on the given devices/ places (in any possible form)?
CHOICE COLUMN - Personal Desktop/Laptop - Official Desktop/Laptop - Pendrive/Portable Hard Disk - Outlook - Email/Data on Cloud (eg. Gmail, Yahoo, Dropbox) - Tablet - Smart phone - DON'T STORE - DON'T HAVE ANY
OPTIONS -

Personal Photographs

Personal Video Files

Personal Audio Files

Bank Statement (Scanned/Digital Copy)

Postpaid Bill (eg. Landline, Electricity Bill etc.) (Scanned/Digital Copy)

Salary Slip (Scanned/Digital Copy)

Air/Railways/Bus Ticket Bookings (Scanned/Digital Copy)

Hotel Booking Confirmation

Sticky notes

Online purchase summary

Insurance/Mutual Fund/Investment/Property Papers (Scanned/Digital Copy)

Admit Card/Mark sheets/Degree (Scanned/Digital Copy)

CV/Resume/Biodata (Scanned/Digital Copy)

Job Offer/Appointment Letter (Scanned/Digital Copy)

Passport (Scanned Copy)

PAN Card (Scanned Copy)

UID Aadhar Card (Scanned Copy)

Credit/Debit/ATM card (Scanned copy/Written details)

License (Scanned Copy)

Voters ID (Scanned Copy)

Medical Reports (Scanned/Digital Copy)

Intellectual Property [your patents, source codes, business plans, novels, poems, research papers, etc.] (Scanned/Digital Copy)

RC of vehicle (Scanned Copy)

Caste Certificate (Scanned Copy)

Income Certificate (Scanned Copy)

Domicile (Scanned Copy)

Marriage Certificate (Scanned Copy)

Birth Certificate (Scanned Copy)

*7. How would you like to rate the following PERSONAL DOCUMENTS on a scale of 1-5, where 1 is for least important and 5 is for most important.
CHOICE COLUMN - 1 - 2 - 3 - 4 - 5 - DON'T HAVE ANY
OPTIONS -

Personal Photographs

Personal Video Files

Personal Audio Files

Visiting Card

Postpaid Bill

Online purchase summary

RC of vehicle

Bank Statement

Salary Slip

Air/Railways/Bus Ticket Bookings

Hotel Booking Confirmation

Sticky notes

Insurance/Mutual Fund/Investment/Property Papers

Admit Card/Mark sheets/Degree

CV/Resume/Biodata

Job Offer/Appointment Letter

Passport

PAN Card

UID Aadhar Card

Credit/Debit/ATM card

License

Voters ID

Medical Reports

Intellectual Property

Caste Certificate

Income Certificate

Domicile

Marriage Certificate

Birth Certificate

Page 3, *Title: Privacy of Stored Data Contd.*

*8. How would you like to rate the following Personally Identifiable Information (PII, i.e the personal information that can be used to uniquely identify you) and other personal information on a scale of 1-5, where 1 is for least important and 5 is for most important.
CHOICE COLUMN - 1 - 2 - 3 - 4 - 5 - N/A
OPTIONS -

Full Name

Father’s Name

Mother’s Maiden Name

Spouse Name

Place of Birth

Parents’ Place of Birth

Gender

Date of Birth

Residential/Office Address

Phone Numbers

Email Address

Passwords

Salary

Bank Details (Acc No. / Credit Card No.)

PAN Number

Passport Number

License Number

Vehicle Number

UID Aadhar Number

ATM PIN Number

Identification Marks

Nicknames

Religion

Caste

Blood Group

Legal Status (Criminal Records)

Biometrics (Fingerprints, Iris Scan)

9. How would you like to rate the following information that you usually store on Online Social Network (OSN, like Facebook, Orkut, Pinterest and Twitter etc.) on a scale of 1-5, where 1 is for least important information (that you can share with everyone) and 5 is for most important information (that you cannot share with anyone).
CHOICE COLUMN - 1 - 2 - 3 - 4 - 5 - N/A
OPTIONS -

Your own pictures

Pictures of you and your spouse

Pictures of you and your children

Pictures of you and your family

Pictures of you and your friends

Pictures of you and your colleagues

Pictures of your spouse only

Pictures of your children only

Pictures of your family only

Pictures of your friends only

Pictures of your colleagues only

Other Pictures clicked by you

Chat Conversations

Status Updates/Tweets

Comments written by you

Activities

Friends List

Followers/Following List

Education Details

Professional Details

Page 4, *Title: Privacy of Stored Data Contd.*

*10. Suppose your digital device (Desktop/ Laptop/ Tablet/ Smart phone) is acquired by the law enforcement/ security agencies for investigating your involvement in a particular case. They get full access to your device and information contained in it. How would this affect your rating (in Question 7) of the importance of PERSONAL DOCUMENTS?

No effect

May increase

May decrease

Other (please specify)

*11. Considering the same scenario as stated in the previous question, how would it affect your rating (in Question 8) of the importance of Personally Identifiable Information (PII) and other PERSONAL INFORMATION?

No effect

May increase

May decrease

Other (please specify)

12. Suppose in the same scenario as stated in the previous question, law enforcement agencies get full access to your online social network account. How would this affect your rating (in Question 9) of the importance of data stored on OSN (Online Social Network)?

No effect

May increase

May decrease

Other (please specify)

*13. Have you ever stored your personal data temporarily on any of your office devices (desktop, laptop etc.) and deleted it after use?

Never

Rarely

Sometimes

Usually

Always

*14. If your digital device (Desktop/ Laptop/ Tablet/ Smart phone) is acquired by the law enforcement/ security agencies for investigating your involvement in a particular case, then do you think the data that you deleted from the device can be recovered?

Yes

No

Maybe

Page 5, *Title: Suggestions and Feedback*

15. It will be great if you give us your valuable feedback (Something that we missed in the list of personal documents, Personally Identifiable Information (PII) and online social network data; or any other comment/suggestion) that would help us to improve our survey.

Page 6, *Title: Demographics*

All the information recorded in this section will be used for research analysis purposes only and in no way can be used to identify you.

*16. Age

18 and under

19-24

25-34

35-44

45-54

55-64

64 and above

*17. Gender

Female

Male

Prefer Not to say

*18. Qualification

Primary School

High School/ Class 10/ Year 10

Intermediate/ Class 12/ Year 12

Undergraduate Diploma

Bachelor’s Degree

Post Graduation

Doctorate

*19. Profession

Academics/ Research

Law Enforcement Agencies

Lawyer

Computer/ IT Industry

Other Engineering Industry

Doctor

Finance/ Banking/ Accounts

Business

Housewife

Others

Other (please specify)

*20. Occupation

Private Sector

Public Sector (Govt. Service)

Business

Unemployed

Student

Others

Other (please specify)

*21. Residential Information -
Drop-down menu for - Country

Other (please specify)

*22. Residential Information -
Drop-down menu for - State

Other (please specify)

23. Which of the following devices do you own/ use?

Tablet

Smart phone

External Hard disks

Laptop

Pen drive

Desktop

Other (please specify)

*24. How long have you been using computers?

0-2 years

2-4 years

4-6 years

More than 6 years

*25. How would you like to rate your experience of computer usage till date? (While rating you shall consider your frequency of computer/ laptop usage per week, knowledge about how computing devices work, proficiency in troubleshooting computer related issues, and awareness about the latest trends in computing devices/ technology)

Basic

Intermediate

Advanced

Expert

NONE OF THE ABOVE

– **End of Survey Questions** –

Appendix B

HACKING CASE QUESTIONNAIRE

1. What operating system was used on the computer?

2. When was the install date?

3. What is the time-zone settings?

4. Who is the registered owner?

5. What is the computer account name?

6. What is the primary domain name?

7. When was the last recorded computer shutdown date/time?

8. How many accounts are recorded (total number)?

9. What is the account name of the user who mostly uses the computer?

10. Who was the last user to logon to the computer?

11. A search for the name of "Greg Schardt" reveals multiple hits. One of these proves that Greg Schardt is Mr. Evil and is also the administrator of this computer. What file is it? What software program does this file relate to?

12. List the network cards used by this computer. This same file reports the IP address and MAC address of the computer. What are they?

13. Find some 'installed programs' that may be used for hacking.

14. What is the SMTP email address for Mr. Evil?

15. List some newsgroups that Mr. Evil has subscribed to?

16. A popular IRC (Internet Relay Chat) program called MIRC was installed. What are the user settings that was shown when the user was online and in a chat channel?

17. This IRC program has the capability to log chat sessions. List some IRC channels that the user of this computer accessed.

18. Ethereal, a popular "sniffing" program that can be used to intercept wired and wireless internet packets was also found to be installed. When TCP packets are collected and re-assembled, the default save directory is that users 'My Documents' directory. What is the name of the file that contains the intercepted data?

19. Viewing the file in a text format reveals much information about who and what was intercepted. What type of wireless computer was the victim (person who had his internet surfing recorded) using?

20. How many files are actually reported to be deleted by the file system?

– **End** –

Bibliography

[1] Jonathon Abbott, Jim Bell, Andrew Clark, Olivier De Vel, and George Mohay. Automated Recognition of Event Scenarios for Digital Forensics. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, pages 293–300, New York, NY, USA, 2006. ACM.

[2] Frank Adelstein. Live Forensics: Diagnosing Your System Without Killing It First. *Communications of ACM*, 49(2):63–66, February 2006.

[3] Ibtesam Al Awadhi, Janet C. Read, Andrew Marrington, and Virginia N. L. Franqueira. Factors Influencing Digital Forensic Investigations: Empirical Evaluation of 12 Years of Dubai Police Cases. *The Journal of Digital Forensics, Security and Law: JDFSL*, 10(4):7, 2015.

[4] Asou Aminnezhad, Ali Dehghantanha, and Mohd. Taufik Abdullah. A Survey on Privacy Issues in Digital Forensics. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, 1(4):311–323, 2012.

[5] Monica Anderson. Smartphone, Computer or Tablet? 36% of Americans Own All Three. `http://www.pewresearch.org/fact-tank/2015/11/25/device-ownership/`, November 2015. Accessed: 2017-06-30.

[6] Daniel Ayers. A Second Generation Computer Forensic Analysis System. *Digital Investigation*, 6:S34 – S42, 2009. The Proceedings of the Ninth Annual DFRWS Conference.

[7] Mridul Sankar Barik, Gaurav Gupta, Shubhro Sinha, Alok Mishra, and Chandan Mazumdar. An Efficient Technique for Enhancing Forensic Capabilities of Ext2 File System. *Digital Investigation*, 4:55 – 61, 2007.

[8] Venansius Baryamureeba and Florence Tushabe. The Enhanced Digital Investigation Process Model. *Digital Investigation*, pages 1–9, 2004.

[9] Nicole Lang Beebe and Jan Guynes Clark. A Hierarchical, Objectives-based Framework for the Digital Investigations Process. *Digital Investigation*, 2(2):147 – 167, 2005.

[10] Nicole Lang Beebe and Jan Guynes Clark. Dealing with Terabyte Data Sets in Digital Investigations. In Mark Pollitt and Sujeet Shenoi, editors, *Advances in Digital Forensics*, pages 3–16, Boston, MA, 2005. Springer US.

[11] Brian D. Carrier. Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers. *International Journal of Digital Evidence*, 1(4):1–12, 2003.

[12] Brian D. Carrier. *File System Forensic Analysis*. Addison-Wesley Professional, 2005.

[13] Brian D. Carrier and Eugene H. Spafford. Getting Physical with the Digital Investigation Process. *International Journal of Digital Evidence*, 2(2):1–20, 2003.

[14] Brian D. Carrier and Eugene H. Spafford. An Event-based Digital Forensic Investigation Framework. In *Digital Forensic Research Workshop*, pages 11–13, 2004.

[15] Brian D. Carrier and Eugene H. Spafford. Automated Digital Evidence Target Definition Using Outlier Analysis and Existing Evidence. In *Digital Forensic Research Workshop*. Citeseer, 2005.

[16] Brian D. Carrier and Eugene H. Spafford. Categories of Digital Investigation Analysis Techniques based on the Computer History Model. *Digital Investigation*, 3:121 – 130, 2006. The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS '06).

[17] Eoghan Casey. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. Academic Press, Inc., Orlando, FL, USA, 3rd edition, 2011.

[18] CFReDS. Hacking Case. `https://www.cfreds.nist.gov/Hacking_Case.html`. Accessed: 2018-02-03.

[19] Séamus Ó Ciardhuáin. An Extended Model of Cybercrime Investigations. *International Journal of Digital Evidence*, 3(1):1–22, 2004.

[20] Frederick B. Cohen. *Digital Forensic Evidence Examination*. Fred Cohen & Associates, 2009.

[21] Neil J. Croft and Martin S. Olivier. Sequenced Release of Privacy-accurate Information in a Forensic Investigation. *Digital Investigation*, 7(1):95 – 101, 2010.

[22] Ali Dehghantanha and Katrin Franke. Privacy-respecting Digital Investigation. In *2014 Twelfth Annual International Conference on Privacy, Security and Trust*, pages 129–138, July 2014.

[23] Sean K. Driscoll. I Messed up Bad: Lessons on the Confrontation Clause from the Annie Dookhan Scandal. *Arizona Law Review*, 56:707, 2014.

[24] Facebook-Business. Finding Simplicity in a Multi-Device World. `https://www.facebook.com/business/news/Finding-simplicity-in-a-multi-device-world`, March 2014. Accessed: 2017-06-30.

[25] Facebook-IQ. The Multidevice Movement: Teens in France and Germany. `https://www.facebook.com/iq/articles/the-multidevice-movement-teens-in-france-and-germany/`, February 2016. Accessed: 2017-06-30.

[26] Simone Fischer-Hübner. *IT-security and Privacy: Design and Use of Privacy-enhancing Security Mechanisms*. Springer-Verlag, Berlin, Heidelberg, 2001.

[27] Organisation for Economic Co-operation and Development. *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. Organisation for Economic Co-operation and Development, 2002.

[28] Simson L. Garfinkel. Digital Forensics Research: The Next 10 Years. *Digital Investigation*, 7:S64 – S73, 2010. The Proceedings of the Tenth Annual DFRWS Conference.

[29] Simson L. Garfinkel. The Expanding World of Digital Forensics. *;login: The Usenix Magazine*, 40(6):12 – 16, December 2015.

[30] Hong Guo, Bo Jin, and Daoli Huang. Research and Review on Computer Forensics. In Xuejia Lai, Dawu Gu, Bo Jin, Yongquan Wang, and Hui Li, editors, *Forensics in Telecommunications, Information, and Multimedia*, pages 224–233. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[31] Md. Mahmud Hossain, Maziar Fotouhi, and Ragib Hasan. Towards an Analysis of Security Issues, Challenges, and Open Problems in the Internet of Things. In *Proceedings of the 2015 IEEE World Congress on Services*, SERVICES '15, pages 21–28, Washington, DC, USA, 2015. IEEE Computer Society.

[32] Shuhui Hou, Tetsutaro Uehara, Siu-Ming Yiu, Lucas C. K. Hui, and Kam-Pui Chow. Privacy Preserving Confidential Forensic Investigation for Shared or Remote Servers. In *2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 378–383, Oct 2011.

[33] Shuhui Hou, Tetsutaro Uehara, SM Yiu, Lucas C. K. Hui, and Kam-Pui Chow. Privacy Preserving Multiple Keyword Search for Confidential Investigation of Remote Forensics. In *2011 Third International Conference on Multimedia Information Networking and Security*, pages 595–599, Nov 2011.

[34] Ricci S.C. Ieong. FORZA – Digital Forensics Investigation Framework that Incorporate Legal Issues. *Digital Investigation*, 3:29 – 36, 2006. The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS '06).

[35] Garda Síochána Inspectorate. Changing Policing in Ireland, 2015.

[36] Umit Karabiyik and Sudhir Aggarwal. Audit: Automated Disk Investigation Toolkit. *The Journal of Digital Forensics, Security and Law: JDFSL*, 9(2):129, 2014.

[37] Simon Key. Flat File Export. `https://www.guidancesoftware.com/app/flat-file-export`. Accessed: 2018-03-10.

[38] Michael Donovan Kohn, Mariki M. Eloff, and Jan H. P. Eloff. Integrated Digital Forensic Process Model. *Computers & Security*, 38:103 – 115, 2013. Cybercrime in the Digital Economy.

[39] Frank Y. W. Law, Patrick P. F. Chan, Siu-Ming Yiu, Kam-Pui Chow, Michael Y. K. Kwan, Hayson K. S. Tse, and Pierre K. Y. Lai. Protecting Digital Data Privacy in Computer Forensic Examination. In *2011 Sixth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering*, pages 1–6, May 2011.

[40] Henry C. Lee, Timothy Palmbach, and Marilyn T. Miller. *Henry Lee's Crime Scene Handbook*. Academic Press, 2001.

[41] David Lillis, Brett Becker, Tadhg O'Sullivan, and Mark Scanlon. Current Challenges and Future Research Areas for Digital Forensic Investigation. *CoRR*, abs/1604.03850, 2016.

[42] Sarah Mocas. Building Theoretical Underpinnings for Digital Forensics Research. *Digital Investigation*, 1(1):61 – 68, 2004.

[43] Adam Moore. Defining Privacy. *Journal of Social Philosophy*, 39(3):411–428, 2008.

[44] Alexios Mylonas, Vasilis Meletiadis, Bill Tsoumas, Lilian Mitrou, and Dimitris Gritzalis. Smartphone Forensics: A Proactive Investigation Scheme for Evidence Acquisition. In Dimitris Gritzalis, Steven Furnell, and Marianthi Theoharidou, editors, *Information Security and Privacy Research*, pages 249–260. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[45] Sebastian Neuner, Martin Mulazzani, Sebastian Schrittwieser, and Edgar Weippl. Gradually Improving the Forensic Process. In *2015 10th International Conference on Availability, Reliability and Security*, pages 404–410, Aug 2015.

[46] Michael G. Noblett, Mark M. Pollitt, and Lawrence A. Presley. Recovering and Examining Computer Forensic Evidence. *Forensic Science Communications*, 2(4), 2000.

[47] Edewede Oriwoh, David Jazani, Gregory Epiphaniou, and Paul Sant. Internet of Things Forensics: Challenges and Approaches. In *9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 608–615, Oct 2013.

[48] Gary Palmer. A Road Map for Digital Forensic Research. In *First Digital Forensic Research Workshop, Utica, New York*, pages 27–30, 2001.

[49] Donn B. Parker. *Crime by Computer*. Scribner New York, 1976.

[50] Pew-Research. Mobile Fact Sheet. *Pew Research Center: Internet, Science & Tech*, January 2017.

[51] Mark M. Pollitt. Computer Forensics: an Approach to Evidence in Cyberspace. In *Proceedings of the National Information Systems Security Conference*, volume 2, pages 487–491, 1995.

[52] Mark M. Pollitt. A Brief History of Computer Forensics. *Unpublished manuscript*, 2004.

[53] Mark M. Pollitt. Six Blind Men from Indostan. In *First Digital Forensic Research Workshop (DFRWS)*, pages 7–8, 2004.

[54] Mark M. Pollitt. A History of Digital Forensics. In Kam-Pui Chow and Sujeet Shenoi, editors, *Advances in Digital Forensics VI*, pages 3–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[55] Darren Quick and Kim-Kwang Raymond Choo. Impacts of Increasing Volume of Digital Forensic Data: A Survey and Future Research Challenges. *Digital Investigation*, 11(4):273 – 294, 2014.

[56] Mark Reith, Clint Carr, and Gregg Gunsch. An Examination of Digital Forensic Models. *International Journal of Digital Evidence*, 1(3):1–12, 2002.

[57] Golden G. Richard III and Vassil Roussev. Next-generation Digital Forensics. *Communications of the ACM*, 49(2):76–80, February 2006.

[58] Marcus K. Rogers. Psychology of Computer Criminals. In *Annual Computer Security Institute Conference, St. Louis, Missouri, USA*, 1999.

[59] Marcus K. Rogers. The Psyche of Cybercriminals: A Psycho-Social Perspective. In Sumit Ghosh and Elliot Turrini, editors, *Cybercrimes: A Multidisciplinary Analysis*, pages 217–235. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[60] Marcus K. Rogers, James Goldman, Rick Mislan, Timothy Wedge, and Steve Debrota. Computer Forensics Field Triage Process Model. *Journal of Digital Forensics, Security and Law*, 1(2):19–38, 2006.

[61] Marcus K. Rogers, Kathryn Seigfried, and Kirti Tidke. Self-reported Computer Criminal Behavior: A Psychological Analysis. *Digital Investigation*, 3:116 – 120, 2006. The Proceedings of the 6th Annual Digital Forensic Research Workshop (DFRWS '06).

[62] Keyun Ruan, Ibrahim Baggili, Joe Carthy, and Tahar Kechadi. Survey on Cloud Forensics and Critical Criteria for Cloud Forensic Capability: A Preliminary Analysis. In *Proceedings of the Conference on Digital Forensics, Security and Law*, pages 55–70, 2011.

[63] Keyun Ruan, Joe Carthy, Tahar Kechadi, and Ibrahim Baggili. Cloud Forensics Definitions and Critical Criteria for Cloud Forensic Capability: An Overview of Survey Results. *Digital Investigation*, 10(1):34 – 43, 2013.

[64] Gong Ruibin, T. Yun, and M. Gaertner. Case-relevance Information Investigation: Binding Computer Intelligence to the Current Computer Forensic Framework. *International Journal of Digital Evidence*, 4(1):147–67, 2005.

[65] Mark Scanlon. Battling the Digital Forensic Backlog through Data Deduplication. In *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 10–14, Aug 2016.

[66] Kimin Seo, Kyungsoo Lim, Jaemin Choi, Kisik Chang, and Sangjin Lee. Detecting Similar Files Based on Hash and Statistical Analysis for Digital Forensic Investigation. In *Proceedings of the 2009 2nd International Conference on Computer Science and Its Applications, CSA 2009*, 12 2009.

[67] Bilal Shebaro and Jedidiah R. Crandall. Privacy-preserving Network Flow Recording. *Digital Investigation*, 8:S90 – S100, 2011. The Proceedings of the Eleventh Annual DFRWS Conference.

[68] Peter Stephenson. Modeling of Post-incident Root Cause Analysis. *International Journal of Digital Evidence*, 2(2):1–16, 2003.

[69] R.B. van Baar, H.M.A. van Beek, and E.J. van Eijk. Digital Forensics as a Service: A Game Changer. *Digital Investigation*, 11:S54 – S62, 2014. Proceedings of the First Annual DFRWS Europe.

[70] Wynand van Staden. Protecting Third Party Privacy in Digital Forensic Investigations. In Gilbert Peterson and Sujeet Shenoi, editors, *Advances in Digital Forensics IX*, pages 19–31. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[71] Robin Verma, Jayaprakash Govindaraj, and Gaurav Gupta. Preserving Dates and Timestamps for Incident Handling in Android Smartphones. In Gilbert Peterson and Sujeet Shenoi, editors, *Advances in Digital Forensics X*, pages 209–225, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[72] Robin Verma, Jayaprakash Govindaraj, and Gaurav Gupta. Data Privacy Perceptions About Digital Forensic Investigations in India. In Gilbert Peterson and Sujeet Shenoi, editors, *Advances in Digital Forensics XII*, pages 25–45, Cham, 2016. Springer International Publishing.

[73] Robin Verma, Anuradha Gupta, Ankit Sarkar, and Gaurav Gupta. Forensically Important Artifacts Resulting from Usage of Cloud Client Services. A Case Study Presented at 2012 Annual Computer Security Applications Conference, Orlando, Florida, USA. `https://www.acsac.org/2012/program/case/Gupta.pdf`, December 2012. Accessed: 2017-06-30.

[74] Svein Yngvar Willassen and S. F. Mjolsnes. Digital Forensic Research. *Telektronikk*, 101(1):92, 2005.

[75] John Zachman. The Zachman Framework for Enterprise Architecture, 2006.