October 2018

# A New Framework for Securing, Extracting and Analyzing Big Forensic Data

Hitesh Sachdev
*Georgia Southern University*

hayden wimmer
*Georgia Southern University*, hwimmer@georgiasouthern.edu

Lei Chen
*Georgia Southern University*

Carl Rebman
*University of San Diego*, carlr@sandiego.edu

# A NEW FRAMEWORK FOR SECURING, EXTRACTING AND ANALYZING BIG FORENSIC DATA

Hitesh Sachdev[1], Hayden Wimmer[1], Lei Chen[1], Carl Rebman[2]
[1]Georgia Southern University
Dept. of Information Technology
P.O. Box 8150
Statesboro, GA 30415, USA
hs02955@georgiasouthern.edu, hwimmer@georgiasouthern.edu*,
lchen@georgiasouthern.edu


[2]University of San Diego
School of Business Administration
5998 Alcala Park
Coronado 212
San Diego, CA 92110
carlr@sandiego.edu

## ABSTRACT

Finding new methods to investigate criminal activities, behaviors, and responsibilities has always been a challenge for forensic research. Advances in big data, technology, and increased capabilities of smartphones has contributed to the demand for modern techniques of examination. Smartphones are ubiquitous, transformative, and have become a goldmine for forensics research. Given the right tools and research methods investigating agencies can help crack almost any illegal activity using smartphones. This paper focuses on conducting forensic analysis in exposing a terrorist or criminal network and introduces a new Big Forensic Data Framework model where different technologies of Hadoop and EnCase software are combined in an effort to promote more effective and efficient processing of the massive Big Forensic Data. The research propositions this model postulates could lead the investigating agencies to the head of the terrorist networks. Results indicate the Big Forensic Data Framework model is capable of processing Big Forensic Data.

**Keywords:** Big forensic data, Forensic analysis, Big Data, EnCase, Hadoop, Map and Reduce, NodeXL, Social Network analysis

# 1. INTRODUCTION

In recent years, advancement in mobile technologies made them an attractive medium for illegal activities. The number of fraud, criminal use and identity theft by mobile and smartphones has dramatically increased (Pascual, Marchini, & Miller, 2018). This demonstrates the need for mobile forensic analysis (Curran, Robinson, Peacocke, & Cassidy, 2012). Smartphones are no longer used to simply connect us with our friends, family and colleagues, but has transformed into a repository which stores all our activities, important dates, numbers, and experiences (audio, video etc.). With internet access commonplace on smartphones, one can send anything at any time. Terrorist and other criminal networks have adopted this powerful leap in technology as a new means of advancing their objectives. From a digital forensics perspective, new methods and frameworks are necessary to process this plethora of gathered information. One such possibility includes combining big data with forensic analysis. This study refers to this combination as Big Forensic Data. Without big data techniques, analyzing the voluminous amount of data stored on smartphones would be nearly impossible or require a massive amount of labor hours.

As the volume, veracity, variety, and velocity of forensic data increases it becomes essential to find methods and techniques to manage and analyze the data and use it for benefit, such as revealing criminal networks. Traditional data warehousing techniques have been inefficient in proactive fraud management and threat detection in large datasets. For example, there is no shortage of multi-billion dollar scams associated with money laundering, bribery, corruption, embezzling, as well as other internet based scams[1] While some of the fraud and attacks were identified and prevented many others were successfully executed. Consequently, traditional analytical techniques could be improved by applying big data techniques for digital forensic analysis.

Big Forensic Data Analysis can help criminal agencies to process information at a higher rate. This paper proposes a framework for the analysis of Big Forensic Data by combining digital forensics techniques with the Hadoop and map reduce framework. The Hadoop framework allows distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up to thousands of machines, each offering local computation and storage (The Apache Software Foundation, 2004) Industry leading forensics software was employed for collecting the information/data from the mobile devices, passing the data through the Hadoop framework to aggregate it, and then applying Social Network Analysis (SNA) on the aggregated data. With the processing power of the Hadoop framework, thousands of devices could be analyzed efficiently thereby reducing the number of labor hours and time to detect criminal activities. Big Forensic Data Analysis can help criminal agencies in the efficient analysis of digital forensic data leading to more timely decision making capabilities. The remainder of this paper is structured as follows: section II presents a review of relevant literature regarding big data and data forensics, section III discusses the Digital Forensic Process, section IV details

---

[1] http://www.fraud.org

the conceptual framework and forensic analysis tools, Section V presents the proposed Big Forensic Data Framework and illustration, section VI presents of the Case study and Hadoop setup and section VII presents the conclusion and the need for additional research.

## 2. LITERATURE SURVEY

### 2.1 BIG DATA

Data is being produced with increasing volume, velocity, and veracity. Five exabytes ($10^{18}$) of data, which was collectively created by humans till 2003 is now produced in just two days. In 2012, the digital world of data expanded to 2.72 zettabytes ($10^{21}$) and is predicted to double every two years (Sagiroglu & Sinanc, 2013). By the year 2020, it is predicted that around 50 billion devices will be connected to the network and the internet (Gerhardt, Griffin, & Klemann, 2012). In 2012, *The Human Face of Big Data* documentary was accomplished as a global project, which centers on real time collection, visualization and analysis of large amount of data. Many statistics were derived from this project. For example, Facebook has 955 million monthly active accounts using 70 languages, 140 billion photos uploaded, 125 billion friend connections. On YouTube, every minute 48 hours of video are uploaded and 4 billion people view videos every day. Google provides many services like monitoring 7.2 billion pages per day, processes 20 petabytes ($10^{15}$) of data daily, and translates the pages into 66 languages. 140 million users send about 1 billion Tweets every 72 hours. 571 new websites are created every minute of the day (Sagiroglu & Sinanc, 2013).

According to the McKinsey Global Institute, the potential of big data lies in five main fields. First, healthcare which includes, clinical decision support systems, individual analytics applied for a patient profile, personalized medicine etc. Second, the public sector which includes creating transparency by accessible related data, discover needs and improve performance. Third, retail which includes product placement design, performance improvement, labor inputs optimization etc. Fourth, manufacturing which includes improved demand forecasting, supply chain planning, sales support etc. Fifth, personal location data which includes smart routing, urban planning, new business models etc.

Sagiroglu and Sinanc (2013) defines big data as massive data sets having a large and complex structure which are difficult to store and analyze. They further describe big data analytics as analysis of the hidden patterns and secret correlations within massive data sets. Labrinidis and Jagadish (2012) developed a panel to discuss the controversies and debunk the myths surrounding big data. The goals of the panel were, to identify how big data is different from the past very large databases and, how data management industries and academia come together to solve big data challenges. Additionally, the panel validated many claims such as Big Data is the same as scalable analytics and that Big Data problems are primarily at the application side.

Big data is gaining popularity and companies like GE and Verizon Wireless are making products that use big data analytics (Davenport, 2014). Big data also promises benefits in healthcare industries in clinical operations, research & development, public health, evidence-based medicine, genomic analytics, pre-adjudication fraud analysis, device/remote monitoring, and patient profile analytics

(Raghupathi & Raghupathi, 2014). Raghupathi and Raghupathi (2014) also discussed the promise and potential of big data analytics in healthcare. Their study describes how big data analytics (BDA) in healthcare is in its initial stage and how the BDA has the potential to transform the way healthcare providers make decisions. Similarly, Patil and Seshadri (2014) discuss how big data is being used in healthcare to lower cost while improving the care process, delivery, and management. They present the state-of-the-art security and privacy issues in big data when applied to the healthcare system.

Marchal, Jiang, State, and Engel (2014) introduced a new architecture for security monitoring of local enterprise networks. The application of this system is to detect and prevent network intrusions. The architecture consists of two systems, one which is dedicated to scalable distributed data storage and management and the second that is dedicated to data exploitation. Katal, Wazid, and Goudar (2013) discussed the importance of Big Data for modern analysis and various issues like Privacy and Security, Data Access, and sharing of information. Guarino (2013) explored the challenges of Big Forensic Data and provides techniques and algorithms that are used in big data analysis that can also be used in digital forensics and mentions techniques like MapReduce and machine learning for the analysis of the forensic disk image as a model to integrate the new paradigm into the established forensic standards.

## 2.2 DIGITAL FORENSICS

According to the Digital Forensic Research Workshop in 2001, Digital Forensic Science is defined by Carrier (2003) as :

*The use of scientifically derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital source for the purpose of facilitating or furthering the reconstruction of events to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.*

Digital data can be obtained from several sources and there are multiple methods and tools to analyze this data. Tassone, Martini, Choo, and Slay (2013) discussed the capabilities of forensic tools for three different platforms: Apple, Android, and RIM's BlackBerry. Tassone et al. (2013) identified where each tool can be best used and provides limitations of tools in accessing different information like call history, message, contacts, media files, and other data. Curran et al. (2012) explored guidelines for maximum data recovery as well as the different types of subscriber identity modules (SIM) available in the market and the different types of analysis that are carried out on the SIM cards for retrieving information. Curran et al, also highlighted the different types of tools available for forensic research at [http://www.mobiledit.com/forensic-solutions](http://www.mobiledit.com/forensic-solutions).

There are several aspects that are considered when aiming to assess the privacy level of an application. The data can exist in various forms: data at rest, data in use, and data in transit. All the different aspects use different methodologies and technologies. Stirparo and Kounelis (2012) demonstrated how different mobile forensics methodologies and tools can be used to assess the privacy level of mobile applications. Catanese, Ferrara, and

Fiumara (2013) presented the LogAnalysis tool that represents and filters visual data, and also statistical analysis features and the possibility of a temporal analysis of mobile phone activities. The tool helps in understanding the structure of the network and the hierarchies of the criminal organization. Ferrara, De Meo, Catanese, and Fiumara (2014), through their work, have tried to achieve two goals: first to provide a conceptual framework for detecting and characterizing criminal organizations from a phone call network. Next, to provide an expert system which helps in unveiling the underlying structure of the criminal network. Grispos, Storer, and Glisson (2011) discussed the different methods and tools to view information held on a Windows mobile device. Their paper seeks to find what information is held on the mobile device. They mainly focus on the use of Celebrite's Universal Forensic Extraction Device (CUFED) as a tool for acquiring data from the mobile phones. They demonstrated that no technique can recover all the information of forensic interest from a mobile device.

Digital forensic tools could be used for analyzing the data from social websites like Facebook, Twitter etc. Al Mutawa, Baggili, and Marrington (2012) use different tools for forensic analysis of this data from mobile devices. Their forensic tests were aimed to recover data from the internal memory of popular smartphones like iPhone, Android, BlackBerry. All smartphones used different tools for forensic analysis: BlackBerry used BlackBerry Desktop Software, iPhone used iTunes application, and unlike the other two devices, Android does not have a management and backup solution. Therefore, the Android phone was rooted using Odin3 which gave access to the protected directories on the system. The results show that nothing could be recovered from BlackBerry devices. On the other hand, iPhone and Android store a significant amount of data which can be used for forensic research.

Digital forensics framework can also help law enforcement agencies. Quick and Choo (2016) developed an approach for data volume reduction which focuses on the registry, documents, spreadsheets, email, internet history, communications, logs, pictures, videos, and other relevant file types. When their approach was applied to the Australian Law Enforcement Agency, the data volume was reduced leaving only the main evidential files and data.

## 3. DIGITAL FORENSIC PROCESS

Computer forensics is a relatively young discipline as compared to other forensic sciences. Because of this, oftentimes the term computer forensics is misunderstood. To address this issue, the Cybercrime Lab in the Computer Crime and Intellectual Property Section (CCIPS) developed a flowchart, which is shown in Figure 2.10, describing the digital forensic analysis methodology. The Cybercrime Lab developed this flowchart after consulting with various computer forensic experts from several federal agencies. It helps in understanding the different elements of the process (Carroll, Stephen K. Brannon, & Song, 2008).

There are three main steps for the analysis of computer forensic data is Data Extraction, Identifications and Analysis represented in Figure 3.10 an outlined in a box.
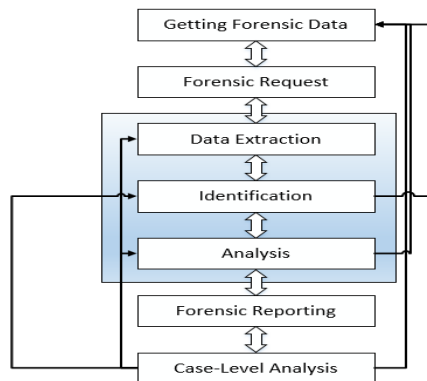
**Figure 3.10: Process Overview adopted from (Carroll et al., 2008)**

### 3.1 DATA EXTRACTION

Examiners start by asking whether they have sufficient information to start the process. If sufficient data is available, they move to the next step where the system is setup and the forensic data is duplicated. Once the forensic data is duplicated, the integrity of the data is analyzed. A plan is developed to extract the data. Examiners should know the data for which they are searching. They add the data into a list known as "Extracted Data List".

### 3.2 IDENTIFICATIONS

Once the data is in the extracted data list, the examiners analyze the type of data. If the data is relevant, then examiners document it into a different document list, the Relevant Data List. If an examiner comes across an item that is incriminating, but outside the scope of the original search warrant, then it is recommended that the examiner immediately stops all activity, notify the appropriate individuals, including the requester, and wait for further instructions. If the data is not relevant the examiners would simply mark it as processed and move on to other relevant elements.

### 3.3 ANALYSIS

In this phase, the examiners try to answers all the questions like who/what, where, when, how. For each relevant item, examiners try to explain when it was created, accessed, modified, received, sent, viewed, deleted, and launched. After the examiner has analyzed the relevant items, they move to the reporting phase in which the examiner tries to document all the finding so that the layman can understand and use them in the case.

## 4. CONCEPTUAL FRAMEWORK

### 4.1 DIGITAL FORENSICS AND BIG DATA

Big Forensic Data and its analysis are an emerging topic and not much work is done in this field. Researchers have tried to combine these two fields but no framework has been designed. Tahir and Iqbal (2015) studied different techniques for the analysis of forensic data from large datasets like Map Reduce, phylogenetic trees, blind source separation, and image culling. Finally, they concluded that Big Forensic Data is heavily dependent on the data source. They also mentioned the factors that should be kept under consideration during the selection of a forensic technique. Zawoad and Hasan (2015) analyzed Big Forensic Data using a map and reduce framework to explore the challenges and issues in forensic paradigm. They concluded that the data is growing very fast, and there is a need for providing support for digital forensic in big data application domain.

## 4.2 DIGITAL FORENSICS ANALYSIS TOOLS

Smartphones are a repository containing huge amounts of information about a user. There are many forensic tools available in the market for processing this data. Table 1 illustrates the live forensic analysis tools which are adapted from Bashir and Khan (2013). In recent years, the market added new and advanced tools which are listed in Table 2 to augment the work by Bashir and Khan (2013) in Table 1.

**Table 1. Live Forensic Analysis Tools adopted from (Bashir & Khan, 2013)**

| Tool Name | Platform | Description |
|---|---|---|
| **AIR (Automated Image and Restore)** | Windows | A Graphical User Interface used to create live image dump of memory. |
| **Autopsy** | Windows/ Linux | This tool is used for disk analysis. |
| **PDUMP** | Windows | This tool is used to take a live memory dump. |
| **WFT (Windows Forensic Toolchest )** | Windows | Used to automate the incident response and perform a live digital analysis. |
| **SMART** | Linux | It is used to perform a live digital forensic analysis. |
| **DEFT (Digital Evidence & Forensic Toolkit)** | Linux | It is used for live computer forensic systems. |
| **BinDiff** | Windows | This tool identifies and highlights the changes and compares the code of program before and after running. |
| **SIFT (SAN Investigation Forensics Toolkit)** | Ubuntu | This tool to perform live digital forensic in the operating system. |
| **IEF (Internet Evidence Finder)** | Windows | It is used to collect evidence from the Internet. |
| **FTK (Forensic Toolkit)** | Windows | This tool is used to the indexing of digital evidence. |
| **DFF (Digital Forensic Framewo rk)** | Windows/ Linux/ Mac | This tool is used for digital evidence collection and analysis toolkit. |

| | Windows | This tool is used to perform forensic analysis on web browsers, emails, Files, and Images. |
|---|---|---|
| **OS Forensics** | Windows | This tool is used to perform forensic analysis on web browsers, emails, Files, and Images. |
| **CAINE (Computer Aided Investigative Environment)** | Linux | For live computer forensics on Linux. |
| **COFEE (Computer Online Forensic Evidence Extractor)** | Windows | It is a toolkit for live digital forensic analysis. |
| **CMAT (Compile Memory Analysis Tool)** | Windows | Memory analyzer. |
| **Wire Shark** | Windows/ Linux/ Mac | Captures and analyzes packets on the network. |
| **Network Miner** | Windows/ Linux | Extracts files, images and other metadata from PCAP files on the network. |

| **Hash Keeper** | Windows | Database application for storing file hash signatures. |
|---|---|---|

**Table 2. New Forensic Analysis Tools**

| Tool Name | Platform | Description |
|---|---|---|
| **EnCase** | Windows | Collect from a wide variety of operating and file systems, including mobile devices(Beneish, Lee, & Tarpley, 2001). |
| **ISafe** | Windows | ISafe is a network and system monitoring digital forensic tool |
| **FTK Imager** | Windows | Allow to acquire images from systems. |
| **ProDiscover** | Windows | Find data on the computer disk. |

### 4.3 ENCASE

Out of the most commonly used computer forensic tools in police departments in the United States is EnCase which is used by about 2000 law-enforcement agencies around the world. It is a 1-mb program written in C++. EnCase can analyze a disk drive from a Windows, Macintosh, Linux, or a DOS machine and make its bit-stream mirror image. To check the authenticity of mirror-image data, EnCase calculates cyclical redundancy

checksums and MD5 hashes (Garber, 2001).

EnCase forensics gives you the ability to quickly search potential evidence to determine whether further investigation is needed. Data could be collected from a wide range of operating and file systems, including mobile devices with EnCase Forensic. EnCase Forensic provides a flexible reporting framework that empowers one to tailor case reports according to specific needs (Encase, 2017; Garber, 2001). William D. Taylor, president and CEO of the International Association of Computer Investigative Specialists says, "It's a great program. I use it every day. It does the work for you if you know what you're doing." (Garber, 2001).

### 4.4    HADOOP

From an open source perspective, big data technologies, like Hadoop, provide cost advantages which is an important factor in any business. Hadoop was created in 2008 by data scientists Doug Cutting and Mike Cafarella. Their aim was to return web search results faster by distributing data and calculations across different computers. Hadoop is an open-source software framework for running very large data sets on clustered environments built from commodity hardware. It provides massive storage, enormous processing power, and also supports hardware failure (Richardson, Tuna, & Wysocki, 2010). Hadoop plays an important role because of its advantages like computing power, distributed storage and distributed processing, which help in faster processing of data. Other features of Hadoop include fault tolerance. Data and application processing are protected against hardware failure. If node fails, Hadoop automatically assigns the job to another node. Most of the databases store the structured data but

Hadoop can also be used for unstructured data. Additionally, the open-source framework is free, and scalable. It can grow by simply adding more nodes (Richardson et al., 2010).

There a few challenges associated with Hadoop. MapReduce programming of Hadoop is not a good match for all problems. There's a widely-acknowledged talent gap. It is hard to find the entry level programmers who have enough java programming skill to use MapReduce. Lastly, Data Security is another challenge associated with big data (Richardson et al., 2010).

### 4.5    NODEXL

NodeXL, Network Overview, Discovery, and Exploration add-in for Excel 2007, 2010, 2013, 2016, adds network analysis and visualization features to spreadsheets. NodeXL takes the data from spreadsheets and converts it into a graph (Smith et al., 2009). NodeXL has many features which make it easy to use represented in Table 3 (MarcSmith, 2016):

### Table 3. Features of NodeXL

| NodeXL | Description |
|---|---|
| **Graph Metric Calculations** | NodeXL can calculate network metrics like degree, closeness centrality, eigenvector centrality, PageRank, clustering coefficient, graph density etc. |
| **Flexible Import and Export Feature** | NodeXL has a predefined format in which it stores the data. NodeXL can import and |

| | export graphs from GraphML, Pajek, UCINet, and matrix formats. |
|---|---|
| **Direct Connections to Social Networks** | NodeXL allows import of network data from Twitter, Facebook, Exchange, Wikis, and YouTube etc. |
| **Zoom and Scale:** | NodeXL allows to zoom into areas of interest, and scale the graph's vertices to reduce clutter. |
| **Easily Adjusted Appearance** | NodeXL allows you to set the color, shape, size, label and much more. |
| **Task Automation** | Using NodeXL can repeat the set of tasks with a single click. NodeXL->Graph->Automate->Run this step executes all that is needed to process a network data set from start to finished, published report |

## 5. BIG FORENSIC DATA FRAMEWORK

Our proposed Big Forensic Data Framework combines all the above-mentioned components, namely digital forensics, Hadoop, and NodeXL. Figure 3.9 represents our system architecture. The different input files illustrated in Figure 3.9 are taken from the digital forensic tool, Encase. These files are passed to the Hadoop framework for processing. The processed data is then fed NodeXL to design a social network graph. This graph helps in exposing a criminal network such as an organized crime or terrorist network.
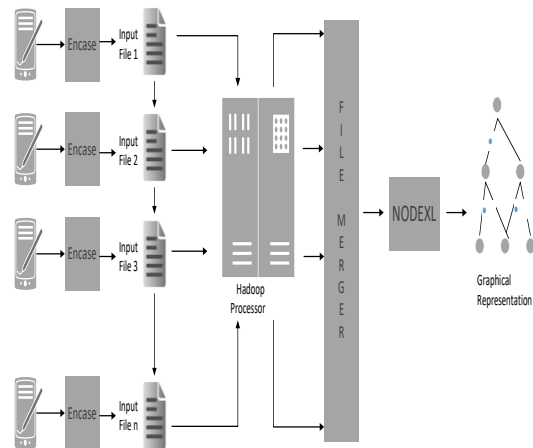


**Figure 3.9: System Architecture**

The Big Forensic Data Framework has different stages:

**Stage 1** represents the smartphones of the convicted criminals that are seized and taken into custody. A smartphone is like a repository in today's modern era and it contains all the information from meeting schedules to personal information that can be obtained about the criminal. This smartphone is passed to the EnCase digital forensic tool for data extraction.

**Stage 2** represents EnCase digital forensic tool which takes the smartphone as an input. This tool makes an image of the device into the local system which can be used by the investigating agencies. In our use case, we extracted the criminal's text message. We can also find all the other information which is available in the smartphone.

Encase generates a report based on the requirement of the investigating agency/officer. This generated report can be seen in Fig 13. It gives all the information like what, when, and where the text message was sent or delivered. This report is unstructured in nature which requires manual human intervention to analyze the report. A separate report is generated for each smartphone. As the report is unstructured in nature and requires a human to perform analysis. We aim to improve this process via employing the Hadoop framework to reduce the amount of human effort required. The Hadoop framework is designed to handle unstructured files. By this process data can be converted into valuable information about the criminal.

The output files from the Hadoop framework are given to file merger process because the NodeXL takes all output files in a specific format. Before passing each file separately to NodeXL, the files are merged and formatted to directly input into NodeXL. Once the output files from the Hadoop framework are changed to the NodeXL format by the file merger it is then passed NodeXL.

NodeXL uses the file generated by file merger to generate the social network analysis (SNA) graph. SNA graph generated by NodeXL can possibly give you the connection of the kingpin with other criminal groups. The graph facilitates examining the network to determine how the network operates and who is connected to whom.

## 6. CASE STUDY

We employed the EnCase forensic tool for extracting the forensic data from mobile devices. The data was extracted from a ZTE phone, model number Z812, android version 5.1 (Lollipop). For extracting the data, a new case was created with the name of "Tablet". Our process follows the standard digital forensic process (Carroll et al., 2008). For step 1, the data extraction phase, we extracted smartphone data from an aforementioned device. Using EnCase we duplicated the smartphone data. For the identification phase, we used the text messages extracted by EnCase. In the final phase, using the SNA graph, we graphed the connections between the text messages. For our experiment, 3 different networks were used. We took one data set from the ZTE phone that represents Hitesh's network. The other two, Max's and Chinmay's networks, were fabricated to support the experiment. In Max's and Chinmay's phones different numbers were used to send text. One common connection, Matt, was used in the phone numbers between all the three different networks.

EnCase has a user friendly graphical user interface (GUI). Clicking on the "acquire smartphone" link as in Figure 6.10, a window will open that will ask about the details of the smartphone. The EnCase tool gives you options for all the different devices like Apple, Blackberry, Nokia, Palm etc. Figure 6.11, on the right side, shows the option of choosing the elements necessary for forensic analysis.
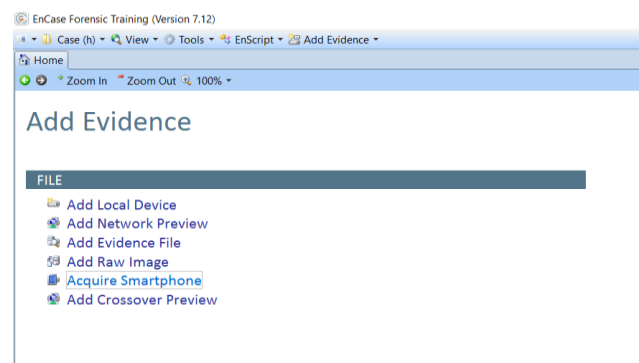


**Figure 6.10: Acquire Smartphone**
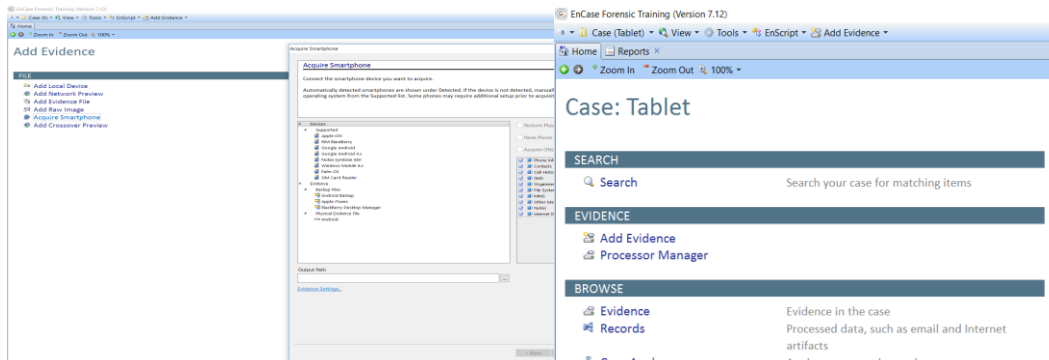
**Figure 6.11: Type of Smartphone**

Figure 6.12 has all the previous cases and is the starting point in the EnCase tool. Figure 6.13 shows when a specific case is read, data is extracted from a device. Figure 6.14 represents the hierarchy of the element of the device. Last, Figure 6.15 shows the extracted report from the EnCase tool. Figure 6.15 gives the option of selecting the elements to include in the final report.
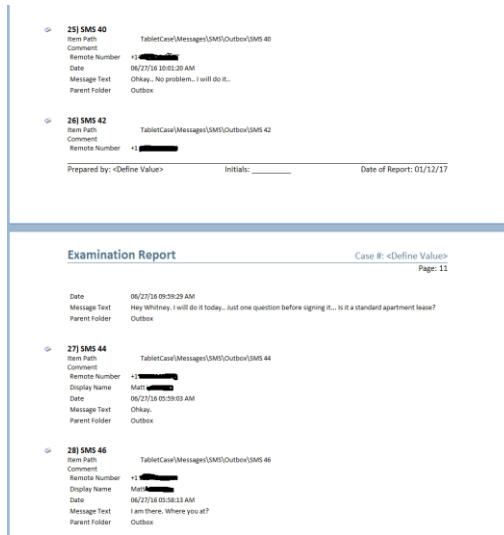


**Figure 6.13: Screen with a specific case**



**Figure 6.12: Home Screen with all the cases**
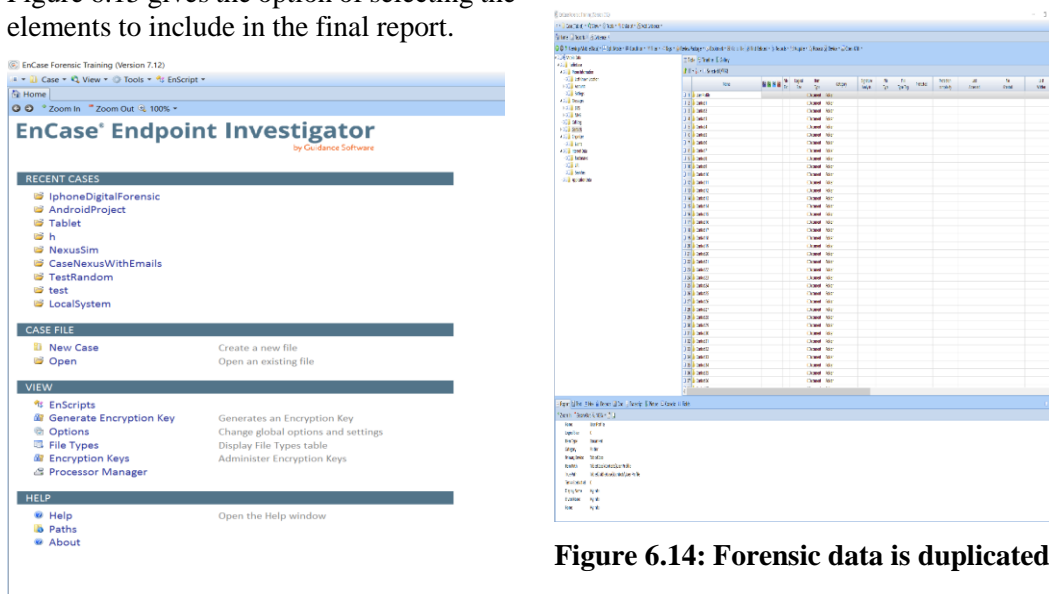


**Figure 6.14: Forensic data is duplicated**

Figure 6.15: Report from EnCase

## 6.1 HADOOP SETUP

The Hadoop Architecture has two main components, the Hadoop Distributed File System (HDFS) and MapReduce (Alam & Ahmed, 2014). HDFS contains three major components Name Node, Data Node, and Secondary Name Node. Name node is also known as the master node. It contains all the information about the data blocks against each data node. Data node is the node which contains the actual data. Upon any request, the data can be read and written to and from this node. The Secondary Name Node is the helper of the Name Node. When name node performs some action on any of the data nodes, it creates a checkpoint and that checkpoint saved on the secondary name node. The system architecture of the HDFS is illustrated in Figure 6.16.
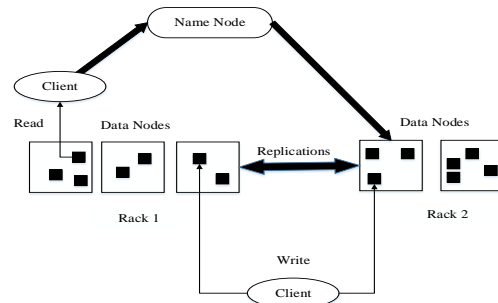


**Figure 6.16: HDFS Architecture adopted from (Borthakur)**

The command "Hadoop fs –put" was used to put the digital forensic report into the HDFS. In our project, we used the command "Hadoop fs –put reportSMS.txt /user/dfProject/BFDFramework/input." To check the text file in HDFS "Hadoop fs –ls" or in our case "Hadoop fs –ls /user/dfProject/BFDFramework/input" was used. To remove a file from the HDFS we use the command "Hadoop fs –rm /user/dfproject/BFDFramework/input/ reportSMS".

MapReduce is a software framework for processing a large amount of data in-parallel on multi-node clustered environment, in a reliable, fault tolerant manner. The task of MapReduce is to divide the data into smaller blocks which are processed individually by the mapper in a parallel manner. Then this data is fed to the reducer. Typically, both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

Figure 3.17 shows the operation of the MapReduce framework.

The command "Hadoop jar BFDFramework.jar org.myorg". BFDFramework "user/dfProject/project1/input /user/dfProject/BFDFramework/output" is

used to execute our MapReduce job. Once the program is successfully executed, the results can be viewed with "Hadoop fs –cat /user/dfProject/BFDFramework/output/*" command.
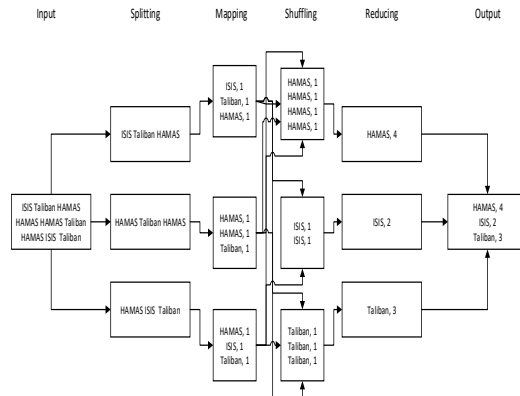


**Figure 6.17: Map Reduce Architecture**

In this work, we created a Hadoop cluster using 3 virtualized machines, one master node and two name nodes that will process our MapReduce job. We created this distributed computing environment for storing and analyzing the huge amount of unstructured data. In our case, the data is reports extracted from the EnCase forensic software. This data is stored on HDFS, separate from a machine's file system. HDFS helps facilitate the distribution of data across the cluster for Hadoop operations.

### 6.2 NODEXL SOCIAL NETWORK ANALYSIS FOR BIG FORENSIC DATA

The output of the Big Forensic Data Framework acts as an input for NodeXL. NodeXL is used for network analysis and visualization. In this paper, we built a case for exposing criminal/terrorist network. After analyzing the data from Hadoop framework, a part of Big Forensic Data, this data is fed to NodeXL. In exposing a

criminal network, we analyzed the text messages sent within the network. After analyzing text messages, we observed that there are 3 different organizations which are illustrated in Figure 6.18, 6.19 and 6.20:
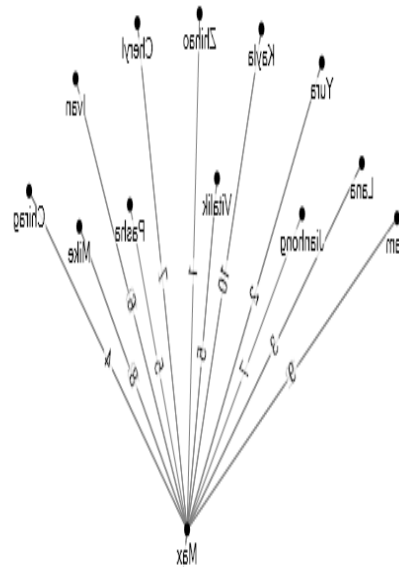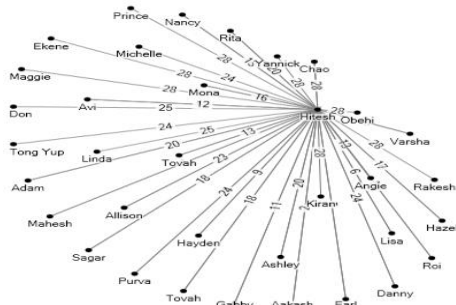


**Figure 6.18 Max's Network**
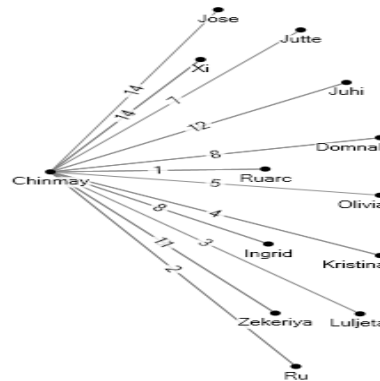


**Figure 6.19: Chinmay's Network**

**Figure 6.20: Hitesh's Network**

When we looked at the bigger picture of the network we observed that Matt was a single link between the heads of all three criminal organizations. This is illustrated in Figure 6.21. Based on this case, an investigator could determine there are 3 connected criminal or terrorist networks. Each network has someone coordinating communications and there is yet another person coordinating between the networks. As is the case with most organized criminal networks, each of the networks, or cells, are unaware of other cells and the larger part of the network and are only concerned with executing their specific tasks. This makes it difficult to extract information from actors in the network; however, employing the Big Forensic Data Framework on seized devices may yield additional information and insight for investigators.
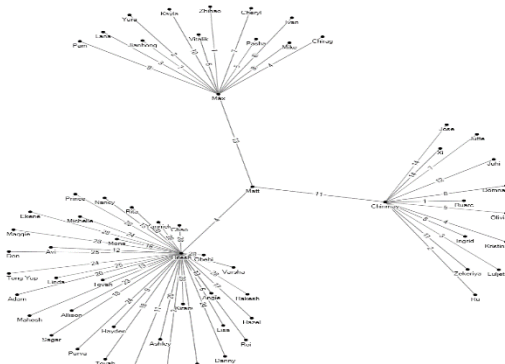


**Figure 6.21. Analysis of Criminal Network**

# 7. CONCLUSION

In today's world terrorism and crime are a major concern. In this work, we focused on conducting forensic analysis to expose a terrorist or criminal network. Criminal and terrorist activities frequently employ mobile technology such as smart phones. Advances in technology facilitate data to grow at exponential rates coining the term Big Data. Similarly, Big Data appears when dealing with forensic analysis of mobile devices coining the term Big Forensic Data. While having this vast amount of data can be beneficial in digital forensic analysis, this data is not useful unless it can be explored appropriately.

The existing analysis tools prove to be inadequate for investigating Big Forensic Data, hence, new methodologies need to be introduced to perform forensic analysis with an expected minimum time response. By combining a different components from both digital forensics, big data, and social network analysis, we designed a new framework to explore Big Forensic Data. We demonstrated how this new Big Forensic Data Framework can be used to expedite and support the investigation process. We also demonstrated that text messages can be used to identify, analyze, and expose terrorist and criminal social networks.

Our paper identifies the challenges of forensic analysis and provides a new approach to digital forensics; however, additional research is needed to solve more challenges in digital forensic analysis and provide opportunities for new insights. We propose a new branch of science with Big Forensic Data and our Big Forensic Data Framework.

# 8. REFERENCES:

Al Mutawa, N., Baggili, I., & Marrington, A. (2012). Forensic analysis of social networking applications on mobile devices. *digital investigation, 9*, S24-S33.

Alam, A., & Ahmed, J. (2014). *Hadoop Architecture and its issues.* Paper presented at the Computational Science and Computational Intelligence (CSCI), 2014 International Conference on.

Bashir, M. S., & Khan, M. (2013). Triage in Live Digital Forensic Analysis. *International journal of Forensic Computer Science, 1*, 35-44.

Beneish, M. D., Lee, C. M. C., & Tarpley, R. L. (2001). Contextual Fundamental Analysis through the Prediction of Extreme Returns. *Review of Accounting Studies, 6*, 165-189.

Borthakur, D. HDFS Architecture Guide. Retrieved from https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Carrier, B. (2003). Defining digital forensic examination and analysis tools using abstraction layers. *International Journal of Digital Evidence, 1*(4), 1-12.

Carroll, O. L., Stephen K. Brannon, & Song, T. (2008). Computer Forensics. *56*.

Catanese, S., Ferrara, E., & Fiumara, G. (2013). Forensic analysis of phone call networks. *Social Network Analysis and Mining, 3*(1), 15-33.

Curran, K., Robinson, A., Peacocke, S., & Cassidy, S. (2012). Mobile phone forensic analysis. *Crime Prevention Technologies and Applications for Advancing Criminal Investigation*, 250.

Davenport, T. (2014). Three big benefits of big data analytics. Retrieved from https://www.sas.com/en_ca/news/sascom/2014q3/Big-data-davenport.html

De Jong, K. A. (2006). *Evolutionary computation : a unified approach*. Cambridge, Mass.: MIT Press.

Encase. (2017). EnCase Forensic Software. Retrieved from https://www.guidancesoftware.com/encase-forensic

Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications, 41*(13), 5733-5750.

Garber, L. (2001). Encase: A case study in computer-forensic technology. *IEEE Computer Magazine January*.

Gerhardt, B., Griffin, K., & Klemann, R. (2012). Unlocking value in the fragmented world of big data analytics. *Cisco Internet Business Solutions Group, June*.

Grispos, G., Storer, T., & Glisson, W. B. (2011). A comparison of forensic evidence recovery techniques for a windows mobile smart phone. *digital investigation, 8*(1), 23-36.

Guarino, A. (2013). Digital forensics as a big data challenge *ISSE 2013 Securing Electronic Business Processes* (pp. 197-203): Springer.

Katal, A., Wazid, M., & Goudar, R. (2013). *Big data: issues, challenges, tools and good practices.* Paper presented at the Contemporary Computing (IC3), 2013 Sixth International Conference on.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with

big data. *Proceedings of the VLDB Endowment, 5*(12), 2032-2033.

Marchal, S., Jiang, X., State, R., & Engel, T. (2014). *A Big Data Architecture for Large Scale Security Monitoring.* Paper presented at the 2014 IEEE International Congress on Big Data.

MarcSmith. (2016, 9/27/2016). NodeXL: Network Overview, Discovery and Exploration for Excel. Retrieved from http://nodexl.codeplex.com/

Pascual, A., Marchini, K., & Miller, S. (2018). *Al Pascual, Kyle Marchini, Sarah Miller*. Retrieved from

Patil, H. K., & Seshadri, R. (2014). *Big data security and privacy issues in healthcare.* Paper presented at the 2014 IEEE international congress on big data.

Quick, D., & Choo, K.-K. R. (2016). Big forensic data reduction: digital forensic images and electronic evidence. *Cluster Computing*, 1-18.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*(1), 1.

Richardson, S., Tuna, I., & Wysocki, P. (2010). Accountinganomalies and fundamental analysis: A review of recent research advances. *Journal of Accounting and Economics, 50*(2-3), 410-454.

Sagiroglu, S., & Sinanc, D. (2013). *Big data: A review.* Paper presented at the Collaboration Technologies and Systems (CTS), 2013 International Conference on.

Smith, M. A., Shneiderman, B., Milic-Frayling, N., Mendes Rodrigues, E., Barash, V., Dunne, C., . . . Gleave, E. (2009). *Analyzing (social media) networks with NodeXL.* Paper presented at the Proceedings of the fourth international conference on Communities and technologies.

Stirparo, P., & Kounelis, I. (2012). *The mobileak project: Forensics methodology for mobile application privacy assessment.* Paper presented at the Internet Technology And Secured Transactions, 2012 International Conference for.

Tahir, S., & Iqbal, W. (2015). *Big Data??? An evolving concern for forensic investigators.* Paper presented at the Anti-Cybercrime (ICACC), 2015 First International Conference on.

Tassone, C., Martini, B., Choo, K.-K. R., & Slay, J. (2013). Mobile device forensics: A snapshot. *Trends and Issues in Crime and Criminal Justice*(460), 1.

The Apache Software Foundation. (2004). What Is Apache Hadoop? Retrieved from http://hadoop.apache.org/

Zawoad, S., & Hasan, R. (2015). *Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities.* Paper presented at the High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on.