IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INTERNET-OF-THINGS (IOT) BIG DATA TRUST MANAGEMENT

Received July 24, 2018, accepted August 24, 2018, date of publication August 28, 2018, date of current version September 21, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2867466

IoT Device Forensics and Data Reduction

DARREN QUICK[1] AND KIM-KWANG RAYMOND CHOO[1,2], (Senior Member, IEEE)
[1]School of Information Technology & Mathematical Sciences, University of South Australia, Adelaide, SA 5001, Australia
[2]Department of Information Systems and Cyber Security, The University of Texas at San Antonio, San Antonio, TX 78249, USA

Corresponding author: Kim-Kwang Raymond Choo (raymond.choo@fulbrightmail.org)

The work of K.-K. R. Choo was supported by the Cloud Technology Endowed Professorship.

ABSTRACT The growth in the prevalence of the plethora of digital devices has resulted in growing volumes of disparate data, with potential relevance to criminal and civil investigations. With the increase in data volume, there is an opportunity to build greater case-related knowledge and discover evidence, with implications at all stages of the digital forensic analysis process. The growth in digital devices will potentially further contribute to the growth in big digital forensic data, with a need for practitioners to consider a wider range of data and devices that may be relevant to an investigation. A process of data reduction by selective imaging and quick analysis, coupled with automated data extraction, gives potential to undertake the analysis of the growing volume of data in a timely manner. In this paper, we outline a process of bulk digital forensic data analysis including disparate device data. We research the process with a research data corpus and apply our process to real-world data. The challenges of the growing volume of devices and data will require forensic practitioners to expand their ability to undertake research into newly developed data structures, and be able to explain this to the court, judge, jury, and investigators.

INDEX TERMS IoT device forensics, data reduction, digital forensics, intelligence analysis.

I. INTRODUCTION

The multitude of interconnected devices include smart vehicles (e.g. unmanned aerial vehicles and autonomous cars), smart fridges, intelligent home assistant devices and systems (e.g. Amazon Echo and Google Home) and Internet of Battlefield / Military Things devices [1], [2]. The volume of data generated from these devices, including Internet of Things (IoT) devices, is significant, and such data can be rapidly transferred from one or more source devices to other connected devices or systems [3]. The data or the systems that stored such data can then be targeted by malicious actors, for example for illicit financial gains (e.g. selling of data exfiltrated from compromised systems). This necessitates the capability to conduct in-depth analysis of the compromised digital devices and systems.

Digital forensics is a process of in-depth analysis of digital devices and data within a legal context, such as a criminal investigation or civil enquiry [4]. The ongoing growth in the number of devices and storage volume requiring analysis puts pressure on timely analysis. The growth in big digital forensic data, which is the large volume, variety, and velocity of data generated by computers and devices has been discussed over many years [5]. The volume of digital forensic data is growing every year, and the variety of data available for forensic analysis is also expanding with the growing popularity and scope of devices and the data they generate [6], [7].

Our recently proposed method of data reduction has demonstrated a capability to reduce the volume of digital forensic data, whilst retaining information in native source file format with original metadata [8], [9]. The data subsets are stored within standard forensic logical (L01) containers, and are able to be processed and examined in a wide variety of digital forensic and analysis software, such as commonly used commercial offerings including EnCase, X-Ways Forensic, NUIX, Magnet Forensic Internet Evidence Finder (IEF), Intella, and Access Data FTK.

Digital forensic subsets are also able to be mounted as logical drives for processing in a wide range of software and analysis tools, such as NetAnalysis, Windows File Analyzer, and RegRipper. This leads us to a situation where data extracted from a wide variety of devices, including mobile phones and cloud stored data, can be collated and amalgamated for analysis purposes.

A process of merging data from a variety of sources can provide benefits to digital forensic analysis in that not only disparate devices, but seemingly disparate cases, may contain linkages which assist to provide a greater understanding of a corpus of data. This may provide breakthroughs, or open avenues of enquiry which may lead to faster resolution of criminal investigations, or reopening of cold case matters with new information.

In this research, we explore the ability to use multiple device and data subsets with reduced storage and processing demands to determine the applicability of cross-device and cross-case analysis with device data, cloud-sourced data, data reduction, and quick analysis techniques.

The contributions of this paper are:

- a process of semi-automated scanning of disparate forensic data subsets, including data from a variety of portable devices, device data, computers, mobile phones, and cloud stored data; and
- cross-device and cross-case analysis leading to greater overall knowledge relating to disparate cases.

In the following paper, digital device data and cross device analysis using data subsets and an automated analysis process is examined. In Section 2, background and related work regarding devices and cross-device analysis are discussed. The process of using Digital Forensic Data Reduction subsets, in conjunction with automated analysis, is then outlined in Section 3 with a view to enable cross-device analysis to be undertaken across multiple device, electronic evidence, and digital forensic data. In Sections 4 and 5, we apply the process of cross device analysis to test data to enable an understanding of its application, and explore the application of the methodology to real-world data, respectively. Sections 6 and 7 summarize the research findings, and conclude the paper and outline future research opportunities.

II. BACKGROUND AND RELATED WORK

Digital devices can range from basic sensors, to sophisticated devices, which can potentially be commandeered for criminal use [10]–[12]. The data on these devices or ''things'' varies according to the device, and can be unstructured, structured, or a combination. Data of potential forensic relevance can be sourced from a variety of devices, with potential implications for digital forensic identification, preservation, collection, analysis, and presentation [1], [10]. These devices present a number of legal and technical issues which include proprietary file and operating systems and communication protocols, encryption, rapid development, and rapid introduction of new devices [13].

Data from a device can be used to prove or disprove information and circumstances, such as the examination of data on a fitness watch, which assisted in solving an investigation into a reported rape [14]. In 2015, a Florida woman was criminally charged after claiming she had been raped whilst staying at her employers' house. Investigators grew concerned of the accuracy of the reported incident and located data on a Fitbit device which indicated she had been moving around during the time she stated she was asleep.

Digital devices have also been used to commit crime, such as using devices for Distributed Denial of Service (DDOS) attacks, commandeering of cloud-based CCTV units, and accessing and using Internet connected printers [15], [16]. It is reported that the LizardStressor malware uses connected digital devices to launch attacks DDOS against banks, telecommunication companies, and government agencies [17]. The malware is successful because of its use of devices which often run embedded Linux based operating systems, have no bandwidth limitations, have minimal security, and often have default passwords shared across devices. Gartner have predicted that by 2020 more than one quarter of cyber-attacks will involve disparate digital devices [17]. They also forecast that 6.4 billion devices will be in use in 2016, an increase of 30% from 2015, and is estimated to reach 11.4 billion devices by 2018. Along with the trade and sale of malware via the Darknet, it is not unrealistic to expect device botnets are currently being traded, along with the availability of live distant video streaming [13].

In other reported incidents, CCTV units commonly referred to as ''nanny cams'' have been accessed and the footage made available to the public [18]. In one incident, hackers accessed the audio functionality and screamed at the occupants children [19]. In another reported instance, ongoing break-ins and nefarious activity included unknown persons moving a camera without being recorded [20]. In another widely reported incident, a hacker accessed thousands of internet connected printers and simultaneously printed out propaganda information, and stated that he had access to more than a million unsecured printers [21]. Whilst the security of these devices is widely discussed [15], there are forensic aspects which also need to be addressed.

Identifying, collecting, interpreting, and presenting the data from disparate digital devices (e.g. in a smart home, Amazon Echo, and other clients devices connected to Amazon Echo, such as mobile devices, smart TV and smart fridge) can be challenging, particularly when data from a variety of devices is merged. In addition, when data from different sources are merged, their relevance to an investigation may become more apparent.

For example, the perennial and often contentious issue, of putting a person at a keyboard, may be answered by a device with biometric information linked to a person of interest wearing or using a device. A wearable device, such as Fitbit wristband, can potentially be used to identify a person via biometric information, such as the heart rate of a wearer. Wearable devices can also have GPS tracking capability. In other words, forensic artifacts obtained from these devices may help to determine if the suspect was at a location of interest on a particular date and time. In the work of Hilts *et al.* [22], for example, a range of fitness trackers were examined from a security aspect. Importantly, the findings also have potential impact in relation to digital forensic analysis, in that it was possible for researchers to perform a man in the middle (MITM) attack on some devices and alter the data in transmission, resulting in false information being pushed into cloud stored data. From this it is possible to conclude that when personal devices such as fitness trackers are sources of evidence or exculpatory information, research will be required by digital forensic practitioners to ensure the data is reliable, as it has been shown it is possible that malicious data

can be inserted, which could be undertaken by the owner or a third party hacker [15].

In addition to issues such as legal issues, privacy, and security concerns [23], we need to consider the resource constraint nature of devices. Specifically, the storage and processing power of many IoT devices can be limited, and hence affect the ability to undertake an investigation. One proposed method to forensically monitor devices is the Forensic Edge Management System (FEMS), which is a proposed device which manages security of smart homes, and collects data of potential forensic value [24].

The increasing number of devices and storage potential is also contributing to a growth in digital forensic evidence, which is overwhelming many digital forensic labs [25]. Current figures indicate the average volume of data in a typical investigation is now close to 3TB, and this is expected to continue to increase [13]. The concerns in relation to the increase in data volume relate to collection and preservation of increasing volumes of data, timely analysis of the increasing volume of data, and storage costs [5]. Data reduction is one proposed method to address collection, preservation, timely analysis, and storage concerns, using a method of Data Reduction by Selective Imaging (DRbSI) [8]. Once data has been collected and stored, there is a need to undertake timely analysis of an increasing volume of disparate data, including the non-standard data generated by devices. This results in a need to be able to undertake rapid processing and analysis of an increasing volume of disparate device and case data.

The process of analysis of a variety of disparate devices is not new to digital forensic analysis. The processes of forensic feature extraction (FFE) and cross device analysis (CDA) were first outlined by Garfinkel [26] and involve extracting information from bulk data, either within a single disk image or across multiple sources. FFE involves scanning a disk or data source for pseudo-unique data identifiers, such as email addresses, email message identifiers, credit card numbers, cookies, and US Social Security numbers. FFE can be used on a single drive to locate information within a disk to speed up initial analysis. Expanding the data identifier points to include disparate personal devices and data may assist in analysis of the growing volume of disparate devices.

Using a process of cross device and cross case analysis has potential to address a range of current issues, which will continue to grow in need and demand. An expanded CDA and FFE with inclusion of specific device identifiers can potentially enable practitioners to discover previously unknown linkages, improve analysis time, and assist with timely analysis. Expanded CDA and FFE also provides for a potential for intelligence analysis across cases and disparate devices. An inhibitor affecting expanded CDA and FFE is the increasing volume of data, the 'big digital forensic data' problem [8]. Using a process of Digital Forensic Data Reduction and Quick Analysis [28], it becomes possible to uncover disparate information linkages across disparate media, including personal devices, and also across disparate cases, in a timely manner.

The process of Quick Analysis enables a practitioner to locate information and intelligence relevant to a case in a timely manner [28]. Using this process across disparate devices, it becomes possible to locate information across multiple devices with a view to merging the findings from disparate data and potentially disparate cases to uncover intelligence and evidence. This can assist with issues of silo-ed investigations, and provide tactical, operational, and strategic intelligence and evidence, to enable decision makers to better understand the context of information.

III. FORENSIC ANALYSIS OF DISPARATE DEVICES AND DIGITAL FORENSIC DATA SUBSETS

The process of digital forensic analysis follows a well-established framework, namely: preparation, identification, preservation, analysis, presentation. The process of preparation includes ensuring agencies and responders have legal authority to act. Impediments can arise when dealing with cloud based data, including legal geographic jurisdiction and privacy aspects. Some jurisdictions can utilize legislative capabilities, such as Australia's *Crimes Act 1914* Section 3L, which provides for warrant holders to access data available to devices at the scene of a warrant search and seizure activity, and Section 3LAA which provides for warrant holders to access data from a device moved from the search location. Section 3LA provides for an order requiring a person to provide information to access data, such as passwords or access codes. The UK (S49 RIPA), India (S69 IT Act), and France (Law #2001-1062 Community Safety) have similar provisions in relation to persons providing passwords, although in the USA, the Fifth Amendment is stated to provide constitutional protection from persons incriminating themselves. In addition, user settings on personal devices may result in data not being stored or accessible and therefore timely action may be required to preserve data, such as cloud stored data or intermediary device stored data.

As more devices become connected to the Internet, uploading data to consumer and corporate cloud storage, the identification of potentially relevant forensic data sources will become even more important. Securing a digital crime scene is problematic, and whilst the physical area is relatively easy to cordon, the wireless crime scene is potentially leaking forensically relevant data whilst crime scene examiners are processing physical devices. Personal devices can be difficult to identify due to the wide range of devices, and the rapid development of technology resulting in devices not previously considered part of a digital crime scene now potentially hosting evidential data, or exculpatory data. As an example, many fridges are now connected devices, with internet browsing capability, data storage, and an ability to notify when items within the fridge are accessed, which could all contribute information to an investigation. Devices may not be discovered until the analysis phase, such as references in web browser history pointing to cloud stored data from a personal device, and it can be difficult to then isolate and seize the appropriate device or data if there is a delay in discovering

a device which may have valuable evidence, resulting in a need for timely collection, processing, and analysis of a large volume of data to determine if other devices are present. Furthermore, even when a device is identified, there are issues surrounding the process of preservation. Do we isolate the device from network connection to prevent data loss, or do we allow devices to continue to collect relevant data, and risk this potentially relevant data being uploaded to cloud storage?

There is also the need to undertake analysis of computers and mobile devices the personal devices have connected or shared data to or with, necessitating a need to collect and preserve a wide range of devices. The quality of the data is also an additional consideration in relation to personal devices, in that some devices function on a ''good enough'' protocol, i.e. as long as enough data is transmitted to form a geocoded timeline or a video stream, there is no need to ensure that every step location or video frame is transmitted and received, and hence there may be dropped data or video frames. Whilst 100% of the data may not be available, the data present may be enough to provide an overview of an event and be suitable upon a test of evidence to be admitted in Court proceedings. Ultimately the decision to admit evidence is in the hands of the trier of facts, but there remains a potential for practitioners to have to test additional aspects of digital forensic data that previously was not required.

Analysis of disparate device data is also an issue to address, in that devices being released on an almost daily basis do not have to comply with forensic readiness principles. Hence, stored data can be in a variety of formats, mostly proprietary; with manufacturers loathe to release information about data structure for fear of competitors copying the formats. With a range of computers and devices seized, there is a need to analyze a range of disparate data from a variety of sources. Malicious actions can also impede an investigation, with potential to sway the findings of an investigation. Many devices are not necessarily NTP time synchronized, and therefore is a possibility for users to manipulate the time and date settings, or even perform a man in the middle attack on wirelessly transferred data to alter information or insert false activity data, such as with the Jawbone UP [22]. The challenge to analyze and determine the authenticity of data will vary according to the device and security measures implemented by the device manufacturers, and with the growing number and volume of devices, will potentially require in-depth research to ensure each device and piece of data is accurate when evidential data is located.

Presentation of disparate device evidential data is also potentially difficult, in that: explaining the relevance, structure, and source of evidence to judge, jury, and the parties involved, when the data structure may have been reverse engineered by the forensic practitioner to enable an understanding of the information, with the research and testing undertaken also required to be presented in a manner which a lay person can understand.

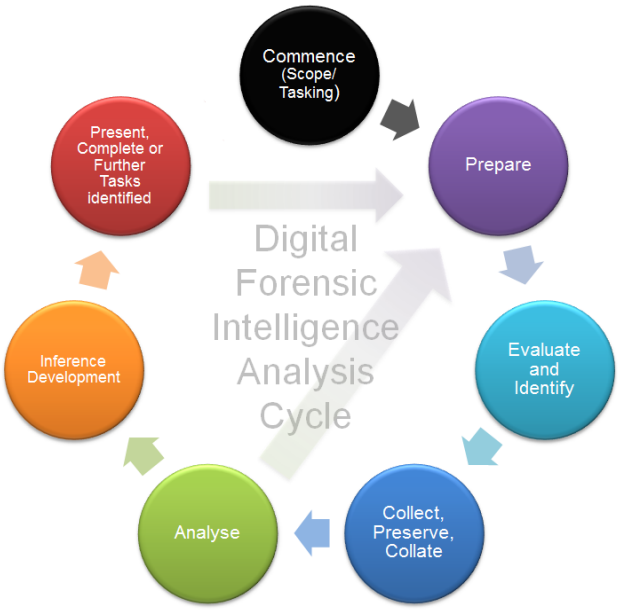

FIGURE 1. Digital forensic intelligence analysis cycle [28].

Using the Digital Forensic Intelligence Analysis Cycle [29] (see Fig. 1) to frame the process, we outline digital forensic disparate device analysis as follows:

- Commence (Scope/Tasking): the aims of the investigation or probe are outlined to a practitioner to enable preparation for the overall examination.
- Prepare: gather the anticipated equipment required and expertise, ensure legal authority exists to act.
- Evaluate and Identify: identify sources of data and potential evidence and intelligence, such as: data which may be on a personal device, mobile phone, computer, storage media, or cloud stored data.
- Collect, Preserve, Collate: once data is identified it needs to be forensically handled, i.e. imaged from a computer or storage device, extracted from a device, preserved from storage media, or a subpoena or legal authority request to a cloud storage provider for cloud stored data, where legal authority exists to collect data. The preservation process can align with a data collection and reduction process, such as that of [8] to collate a subset of relevant data.
- Analyze: data subsets and data extracted from computers, mobile phones, devices, including data from cloud stored providers, is then examined for evidence or intelligence, with a focus of that of the scope of the task [28]. If during the process of analysis, additional sources of data are identified, such as another device, media storage, or cloud storage, the process forks back to preparation to collect new data, whilst analysis progresses.
- Inference Development: with the knowledge gained during the process of collation and analysis, ideas are formed in relation to the questions of who, how, what, when, why, and where. The gained knowledge is used to

form inferences about the investigation or intelligence probe to answer questions or outline findings in relation to evidence and intelligence.

- Present, Complete, or Further Tasks Identified: the findings of the overall process of analysis are formed into a report (written and/or verbal) which is communicated to the requesting persons involved in the legal process or probe. If further tasks are identified, the process continues in the cycle until complete. Feedback is provided to relevant parties, and also sought to ensure the goals of the investigation have been achieved.

IV. PERSONAL DEVICE AND TEST DATA ANALYSIS

As discussed in Sections 2 and 3, personal devices are increasingly prevalent and becoming more frequent in digital forensic investigations. This results in a larger variety of disparate data, with associated issues relating to data source, volume, and type. Processes to undertake digital forensic preparation, identification, preservation, analysis, and reporting are required to include the additional scope and focus of disparate device investigations, including methods to undertake analysis of devices and data in conjunction with computer data and cloud stored data, in a timely manner. The focus of this research is on two aspects; (a) to explore the reliability of data from a fitness band device, and (b) to examine a process of analysis of a large volume of disparate data, including personal device data, mobile phone extracts, operating system images, storage media, and cloud stored data collections.

The software used for this research included; Bulk Extractor, NUIX, EnCase, RegRipper, IEF, NetAnalysis, and Pajek64. Bulk Extractor is software that scans a disk image or directory of files and extracts information such as credit card numbers, email addresses, web addresses, and telephone numbers (www.forensicwiki.org/wiki/bulk_extractor). NUIX is software which scans through a wide range of data and processes this to extract information and enable analysis (www.nuix.com). EnCase is forensic software which provides for analysis of forensic images, data, and files (www.guidancesoftware.com). RegRipper is used to scan Windows Registry Files and output the extracted information (www. forensicwiki.org/wiki/regripper). IEF is used to scan forensic images and data for a range of information including internet history, chat history, and operating system information (www.magnetforensics.com/magnet-ief). NetAnalysis is used to scan forensic images and mounted volumes for internet history and output the information in a spreadsheet format (www.digital-detective.net/digital-forensic-software/netanalysis). These tools are commonly used for digital forensic analysis. Pajek64 is software which enables the charting of large network data and entity information, such as in social network analysis. In this research is used to prepare entity interlink link charts as used in intelligence analysis (www.mrvar.fdv.uni-lj.si/pajek/).

The data from fitness bands has been crucial in investigations, such as Snyder [14] in which the device data showed that the allegations were false as the user was moving around

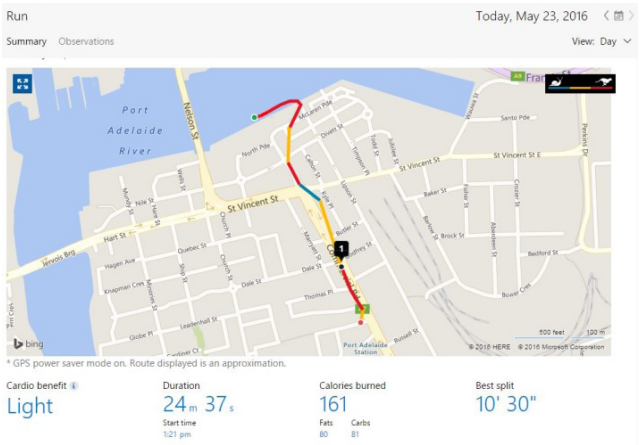

FIGURE 2. GPS data mapped from Microsoft Band 2 including corrected start time of session.

at the time when it was stated she was asleep. As discussed in the previous sections, there is also a potential for users to alter or manipulate data from fitness bands. Hence, there may be a need to examine data available from personal devices such as fitness bands in conjunction with the data from other devices, such as portable storage, mobile phones, computers, and cloud stored data, to determine the actuality an event or events, and explore the evidential or intelligence data in a wider context of events.

In the first aspect of this research we explore the process of altering the time settings on a fitness band device, in conjunction with the analysis of personal fitness devices along with disparate case data holdings. A Microsoft Band 2[1] device was available for research. For research purposes, the device time was intentionally set to a different time in an endeavor to provide a false alibi relating to the time of recorded activity. A short walk was undertaken with the device time set earlier than current time, and then synced and the data uploaded to the associated cloud stored account. Using the device web interface, we exported the GPS data from the web application. We observed that when uploaded, the start time and the waypoint times were corrected to the actual time, both in web presented data and in the exported GPS data, refuting the attempt to create a false alibi. Fig. 2 displays the corrected time for the GPS coordinates from the web browser.

The second aspect of this research is to test a process as would be seen in real world investigations involving a wider range of disparate device data requiring bulk data analysis in a timely manner. We explore a process of semi-automated scanning of multiple forensic data subsets using Bulk Extractor software [27].

To undertake research in relation to the growth in the number of disparate devices and volume of storage media typically encountered in an investigation, which can be in a variety of differing formats, we used test data from the M57 corpus [30] as this provides for a large but manageable

[1]Researchers did not have access to a Jawbone UP as per (every step you fake) but did have access to a Microsoft Band 2 device for testing.

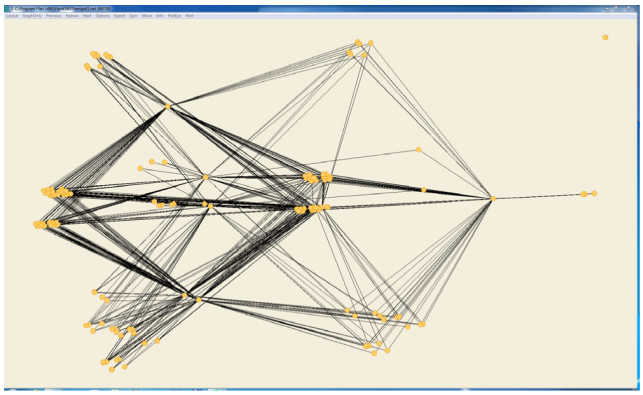

FIGURE 3. Pajek entity link chart from M57 corpus.

volume of data from a variety of devices such as would be seen in a typical investigation. The M57 corpus is a collection of a variety of user cases containing forensic computer images and mobile device extracts in a variety of formats. The M57 corpus is made available for the purposes of research. We previously used this data for data reduction research, where we demonstrated the ability to reduce total data volume using the Data Reduction by Selective Imaging (DRbSI) method [8] and a process of Quick Analysis [28] to distil information relevant to the task or scope of analysis. We reduced the forensic data from the M57 computers, portable storage devices, mobile phones, and tablet devices, with source data comprising approximately 498GB, which reduced to 4.25GB of extracts and forensic container files encompassing potentially relevant data.

For this research, Bulk Extractor v1.5.5 was used to scan the test data subsets, resulting in 2.02 GB of output, comprising 23,496 email features, and 22,962 picture files, in approximately 30 minutes. The mobile phone spreadsheet reports from 41 mobile devices, comprising 207 MB, was previously merged into one text file and converted to Pajek64 format for analysis [29]. The data output from undertaking Quick Analysis [28] of the subsets, including RegRipper Registry extracts, IEF spreadsheet reports, NetAnalysis csv output, and data extracted from other sources within the DRbSI subsets, was merged with the output from Bulk Extractor, along with the previously merged mobile phone extracted data, resulting in a large file of extracted information with associated source and relationship links. This data was then loaded into Pajek64. The resulting entity link chart (see Fig. 3) displays commonality linkages between 443,589 entities within the extracted data.

We merged the Fitness band data with the Pajek data and overall case data, linking the GPS coordinates with the entity associated with the device, along with data extracted from a mobile phone and computer, also associated with the entity. Using this data to create a spreadsheet timeline of events from the overall data enabled analysis of events in context, using the merged data from the computers, mobile phones, tablets, DRbSI subsets, Quick Analysis data, and other associated extracted data. The process of merging the disparate information demonstrated an ability to undertake analysis of a wide range of unrelated data extracted from disparate devices and sources, using the time and date as a common point of reference to align the data.

Further analysis of the data and linkages would then be undertaken, including links with multiple connections, and outliers, with a focus on the scope or aim of the investigation. Data from a wide range of systems can be extracted, collated, and merged with other disparate data for analysis, with entity, interlink, and timeline analysis methodologies appropriate to gain an understanding of the scope of the data holdings. Further data may be available from security systems, such as smart door locks which record biometric information as a person enters or exits a smart home, or a wireless internet connected doorbell system with video which records movement of persons near the device.

This process highlights an ability to rapidly process a large volume of test data from a range of disparate devices, and the potential to locate further devices and data which may be relevant to an investigation, such as personal fitness device device data. Whilst the test data corpus is reflective of real world investigations, there is still a need to examine the potential for the process to be applied to real world investigation data.

V. ANALYSIS OF REAL WORLD DATA

To undertake research into automated extraction of information from real world big digital forensic data, we were afforded limited access to South Australia Police Electronic Evidence Section historic data backups. We did not examine the contents of case data, and only reviewed the time to undertake processing of limited historic meta-data. The software used for this research again included NUIX, EnCase, RegRipper, IEF, NetAnalysis, and Bulk Extractor.

Europol reports that the average volume of data per investigation is now close to 3TB, and this can include data from a wide range of portable devices, mobile phones, computers, and cloud stored data [13]. A real-world historical case which comprised of 18 computers, laptops, portable storage, mobile phones, and tablet devices totaling 2.7 TB of source data, potentially containing disparate device data, such as personal fitness devices or other disparate devices syncing data with mobile phones of computers. In this matter, full imaging reportedly took approximately 42 hours, and the forensic image files were loaded into NUIX 6.2.3, and took 65 hours to process the full forensic images, resulting in a total of 107 hours before the evidence was ready for review by an investigator. When the DRbSI process was applied to the forensic image data, according to the process in Quick and Choo [8], the process took less than 4 hours to collect 46.1GB of DRbSI subsets, and 9 hours to process the DRbSI subsets in NUIX, a total of 13 hours.

To test the use of a semi-automated analysis process on real world data, the forensic image source data was processed with Bulk Extractor software, which took 43 hours and 11 minutes to complete. In comparison, the DRbSI subsets were also processed with Bulk Extractor, processing in 1 hour 19 minutes.

Due to the nature of the data and the limited access provided, the actual data itself was not able to be viewed or queried in an effort to determine the volume of disparate device data contained within the data, and this remains a future research opportunity should further access be provided to real world data. Hence this research focused on the application of the process to a large volume of data and the associated time savings rather than the contents of the data.

In a further effort to explore the timeliness of processing big digital forensic data, the DRbSI subsets from 544 devices was loaded into NUIX 6.2.3, EnCase 6.19.7, and EnCase 7.10.5. Again, the data was not viewed, rather, the times for processing was noted. EnCase 6.19.7 took approximately three (3) minutes to load and open the 544 L01 files. File signature analysis was run, and took 2 hours and 8 minutes. Over 10 million files were presented for analysis, including 907,015 documents, 52,742 emails, 2,221,521 picture files, and 2,333 container files. Within this data was potentially relevant disparate device data.

In comparison, EnCase 7.10.5 took 27 hours to open the L01 files. Signature analysis took 6 hours 15 minutes, and identified 37,224 emails. The L01 data from the 544 files was also loaded into NUIX 6.2.3, and some simple metadata analysis was able to be conducted, such as identifying device type; such as iPhone and Samsung mobile phones, Panasonic, Nikon, and Canon camera identifiers in picture EXIF data, and nearly 3,000 resume documents (further analysis was not undertaken on this data).

This research demonstrated a capability to load DRbSI subset L01 files from a large number of disparate devices ranging from portable storage to multi-terabyte hard drives, and it was possible to load and process these with EnCase 6.19.7, EnCase 7.10.5 and NUIX 6.2.3, and potentially conduct in-depth analysis of the data, along with using bulk_extractor across the subsets. The data encompassed within the disparate devices potentially contained fitness device data, but in-depth analysis of this information was not undertaken.

VI. DISCUSSION

As personal devices, computers, portable storage, mobile phones, and tablets become more pervasive throughout society, there will be a growing need for forensic analysis of these devices. As these devices store data in a variety of formats, including sending the data from the device to a connected mobile phone, tablet, to cloud storage, and then viewing on a computer, there is a growing need for digital forensic practitioners to be able to identify, collect, analyze, and present the data from these devices in a manner that a legal environment can understand the implication of the evidence, and in a timely manner. The ability of users and malicious actors to manipulate the data is also an issue that needs to be considered when evaluating data and information for evidential and intelligence potential. In our testing undertaken with a fitness band we were able to show that a malicious user attempting to alter the time of activity to provide a false alibi was unable to be completed as the time and date settings were corrected when uploaded to the associated account. However, a determined user could acquire a different device, such as the Jawbone UP, and inject false information into the data with a MITM attack [22].

With the growing volume of disparate data, there is a need to be able to undertake analysis on growing volumes of structured and unstructured data. The method outlined in the previous sections as applied to test data (M57) and real world data, demonstrated an ability to undertake analysis of a large volume of disparate data, and locate potential evidence and intelligence in a timely manner. Using Bulk Extractor, it was possible to scan DRbSI subsets in a semi-automated manner, and then merge the output to analyze a large volume of data in a timely manner for linkages across devices and cases. Loading real world DRbSI subsets into common digital forensic software (EnCase and NUIX), it was possible to undertake analysis across a large volume of devices in a much shorter timeframe. Future research potential includes analysis of devices to locate data which can assist with entity extraction, and include this information in bulk extractor search configuration files. Machine learning, as used with big data analytics, could also be explored for potential use with big digital forensic data.

As more and more devices are seized and presented for digital forensic analysis, there will be a larger source of data for analysis, potentially locating evidence and intelligence to enable investigators and decision makers a greater understanding of events from large volumes of data. Strategic and management level information can be gleaned from data; operational knowledge can be located and provided to investigators and managers, including information relating to crime trends. Tactical, target specific information can also be located and communicated in a timely manner.

In the growing digital age forensic practitioners will need to shift the focus from computers and mobile phones, and will need to focus on a range of potential data sources, including personal devices, portable devices, and cloud storage. There will be a need to be able to collect and preserve large volumes of data from a range of devices and cloud stored data, i.e. whichever data enables a decision to be made. This has to be balanced with the growing volume of data, and the need for timely analysis, hence a process of data reduction and bulk data analysis will become more necessary in the coming years.

Digital forensic practitioners will need to be able to identify disparate sources of evidence, test methods of extracting data in a repeatable manner, and undertake research to determine the validity of the extracted data. As devices become more prolific, it will become necessary to develop these skills, and work in conjunction with academic and commercial entities to determine best practice for analysis and understanding of disparate data. In future, there will be potentially many devices for academics and commercial entities to provide forensic extract and analysis solutions; hence, collaboration may assist in obtaining greater overall results.

VII. CONCLUSION

As connected devices become more pervasive and prevalent throughout society, there will be an impact on demand for digital forensic analysis of these devices, and the data they generate, wherever it may be stored. A range of additional considerations for practitioners include identification, collection, preservation, analysis, authentication, and presentation of disparate device data. First responders at a crime scene will need to consider personal and Internet connected devices within and external to the crime scene.

Along with the need to secure the physical crime scene, there will be a need to potentially secure the wireless crime scene, as potentially data may be leaking whilst crime scene examiners are processing physical devices. Considerations will now need to include: who, how, what, when, where, why, and which device is relevant, and also the authenticity of the data. Along with the growing volume of disparate device data, there is also a growth in computer, portable storage, and mobile phone device data. Digital forensic analysis in future will need to be able to examine a wide range of data and devices in a timely manner.

Digital forensic practitioners will need to focus on relevant data, which may not necessarily be on a device. The relevant data may be on a personal fitness device, transferred to a mobile phone, computer, or uploaded to cloud storage. There is a need to gather data from a variety of sources and undertake rapid analysis on a range of data structures, which assist in evidence and intelligence identification, in a timely manner.

As demonstrated in this research, using a process of data reduction and semi-automated subset analysis with Bulk Extractor software, charted with entity linking software such as Pajek64, enables timely analysis of a wide range of disparate data. Future research opportunities include the potential for supervised and unsupervised learning algorithms, as used in big data analytics, to be adapted for use with big digital forensic data. The scope of Bulk Extractor can also be expanded with other device specific data included in the process of identification and entity extraction.

REFERENCES

[1] E. Oriwoh, D. Jazani, G. Epiphaniou, and P. Sant, "Internet of Things forensics: Challenges and approaches," in *Proc. 9th Int. Conf. Collaborative Comput., Netw., Appl. Worksharing (Collaboratecom)*, 2013, pp. 608–615.

[2] A. Cassidy. (May 3, 2016). *The 'Internet of Things' Revolution and Digital Forensics*. [Online]. Available: http://www.nuix.com/2014/02/19/the-internet-of-things-revolution-and-digital-forensics

[3] J. Huang. (Apr. 27, 2016). *Extracting My Data from the Microsoft Band*. [Online]. Available: http://jeffhuang.com/extracting_my_data_from_the_microsoft_band.html

[4] D. Quick, B. Martini, and K.-K. R. Choo, *Cloud Storage Forensics*. Amsterdam, The Netherlands: Elsevier, 2014.

[5] D. Quick and K.-K. R. Choo, "Impacts of increasing volume of digital forensic data: A survey and future research challenges," *Digital Invest.*, vol. 11, no. 4, pp. 273–294, 2014.

[6] N. D. W. Cahyani, B. Martini, K.-K. R. Choo, and A. K. B. P. M. N. Al-Azhar, "Forensic data acquisition from cloud-of-things devices: Windows smartphones as a case study," *Concurrency Comput., Pract. Exper.*, vol. 29, no. 14, p. e3855, 2016.

[7] Q. Do, B. Martini, and K.-K. R. Choo, "Is the data on your wearable device secure? An Android Wear smartwatch case study," *Softw., Pract. Exper.*, vol. 47, no. 3, pp. 391–403, 2017.

[8] D. Quick and K.-K. R. Choo, "Big forensic data reduction: Digital forensic images and electronic evidence," *Cluster Comput.*, vol. 19, no. 2, pp. 723–740, 2016.

[9] D. Quick and K.-K. R. Choo, "Data reduction and data mining framework for digital forensic evidence: Storage, intelligence, review and archive," *Trends Issues Crime Criminal Justice*, vol. 480, pp. 1–11, Sep. 2014.

[10] R. Hegarty, D. Lamb, and A. Attwood, "Digital evidence challenges in the Internet of Things," in *Proc. 10th Int. Netw. Conf. (INC)*, 2014, p. 163.

[11] C. J. D'Orazio, and K.-K. R. Choo, "A technique to circumvent SSL/TLS validations on iOS devices," *Future Gener. Comput. Syst.*, vol. 74, pp. 366–374, Sep. 2017.

[12] C. J. D'Orazio, K.-K. R. Choo, and L. T. Yang, "Data exfiltration from Internet of Things devices: iOS devices as case studies," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 524–535, Apr. 2017.

[13] *The 2016 Internet Organised Crime Threat Assessment (IOCTA)*, Europol, Eur. Law Enforcement Agency, The Hague, Netherlands, 2016.

[14] M. Snyder. (Apr. 27, 2016). *Police: Woman's Fitness Watch Disproved Rape Report*. [Online]. Available: http://abc27.com/2015/06/19/police-womans-fitness-watch-disproved-rape-report/

[15] Q. Do, B. Martini, and K.-K. R. Choo, "A data exfiltration and remote exploitation attack on consumer 3D printers," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2174–2186, Oct. 2016.

[16] O. Osanaiye, K.-K. R. Choo, and M. Dlodlo, "Distributed denial of service (DDoS) resilience in cloud: Review and conceptual cloud DDoS mitigation framework," *J. Netw. Comput. Appl.*, vol. 67, pp. 147–165, May 2016.

[17] W. Ashford. (Oct. 9, 2016). *LizardStresser IoT botnet Launches 400 Gbps DDoS Attack*. [Online]. Available: http://www.computerweekly.com/news/450299445/LizardStresser-IoT-botnet-launches-400Gbps-DDoS-attack

[18] K. Hill. (Oct. 9, 2016). *Watch Out, New Parents—Internet-Connected Baby Monitors are Easy to Hack*. [Online]. Available: http://fusion.net/story/192189/internet-connected-baby-monitors-trivial-to-hack/

[19] K. Hill. (Oct. 9, 2016). *The Terrifying Search Engine that Finds Internet-Connected Cameras, Traffic Lights, Medical Devices, Baby Monitors and Power Plants*. [Online]. Available: http://www.forbes.com/sites/kashmirhill/2013/09/04/shodan-terrifying-search-engine/#6676a2b1174c

[20] M. Harthorne. (Oct. 9, 2016). *Couple Believes Burglar Hacked Into Nanny Cam to Spy on Them*. [Online]. Available: http://komonews.com/news/local/couple-believes-burglar-hacked-into-nanny-cam-to-spy-on-them

[21] J. Wieczner. (Oct. 9, 2016). *This Hacker Sent Nazi Flyers to Thousands of Printers in Internet of Things 'Experiment*. [Online]. Available: http://fortune.com/2016/03/29/hack-printers-internet-of-things/

[22] A. Hilts, C. Parsons, and J. Knockel. (Oct. 9, 2016). *Every Step You Fake: A Comparative Analysis of Fitness Tracker Privacy and Security*. [Online]. Available: https://openeffect.ca/fitness-trackers/

[23] E. Oriwoh, H. Al-Khateeb, and M. Conrad, "Responsibility and non-repudiation in resource-constrained Internet of Things scenarios," in *Proc. Int. Conf. Comput. Technol. Innov. (CTI)*, 2016.

[24] E. Oriwoh and G. Williams, "Internet of Things: The argument for smart forensics," in *Proc. Handbook Res. Digit. Crime, Cyberspace Secur., Inf. Assurance*, 2014, pp. 407–423.

[25] H. Parsonage. (Aug. 4, 2013). *Computer Forensics Case Assessment and Triage—Some Ideas for Discussion*. [Online]. Available: http://computerforensics.parsonage.co.uk/triage/triage.htm

[26] S. L. Garfinkel, "Forensic feature extraction and cross-drive analysis," *Digit. Invest.*, vol. 3, pp. 71–81, Sep. 2006.

[27] S. Garfinkel, "Digital media triage with bulk data analysis and *bulk_extractor*," *Comput. Secur.*, vol. 32, pp. 56–72, Feb. 2013.

[28] D. Quick and K.-K. R. Choo, "Big forensic data management in heterogeneous distributed systems in smart cities: Quick analysis of multimedia forensic data," *Softw., Pract. Exper.*, vol. 47, no. 8, pp. 1095–1109, 2016.

[29] D. Quick and K.-K. R. Choo, "Pervasive social networking forensics: Intelligence and evidence from mobile device extracts," *J. Netw. Comput. Appl.*, vol. 86, pp. 24–33, May 2017.

[30] S. Garfinkel, R. Farrell, V. Roussev, and G. Dinolt, Bringing science to digital forensics with standardized forensic corpora. DFRWS, Montreal, QC, Canada, 2009. Accessed: Sep. 9, 2013. [Online]. Available: http://digitalcorpora.org/corpora/disk-images

DARREN QUICK received the Ph.D. degree in computer and information science from the University of South Australia, Australia, in 2017. He is a Senior Intelligence Technologist with the Australian Department of Home Affairs and a former Digital Forensic Investigator with the Australian Border Force, and previously an Electronic Evidence Specialist with the South Australia Police. He has undertaken over 650 digital forensic investigations involving many thousands of digital evidence items. In 2012, he was awarded membership of the Golden Key International Honour Society; in 2014, he received a Highly Commended Award from the Australian National Institute of Forensic Science; and in 2015, he received the Publication of the Year Award from the Australian Institute of Professional Intelligence Officers.

KIM-KWANG RAYMOND CHOO (SM'15) received the Ph.D. degree in information security from the Queensland University of Technology, Australia, in 2006. He currently holds the Cloud Technology Endowed Professorship with The University of Texas at San Antonio (UTSA), and has a courtesy appointment at the University of South Australia. In 2016, he was named the Cybersecurity Educator of the Year - APAC (Cybersecurity Excellence Awards are produced in cooperation with the Information Security Community on LinkedIn), and in 2015, he and his team won the Digital Forensics Research Challenge organized by University of Erlangen-Nuremberg, Germany. He is also a fellow of the Australian Computer Society and an Honorary Commander of the 502nd Air Base Wing, Joint Base San Antonio-Fort Sam Houston. He was a recipient of the 2018 UTSA College of Business Col. Jean Piccione and Lt. Col. Philip Piccione Endowed Research Award for Tenured Faculty, the IEEE TrustCom 2018 Best Paper Award, the ESORICS 2015 Best Research Paper Award, the 2014 Highly Commended Award by the Australia New Zealand Policing Advisory Agency, the Fulbright Scholarship in 2009, the 2008 Australia Day Achievement Medallion, and the British Computer Society's Wilkes Award in 2008.