Abstract

As digital storage increases exponentially in size and becomes cheaper and more easily available, the volume of digital artefacts of evidential value increases with it. The current convention of digital forensics is that evidence is extracted, analysed and reported in relation to a single investigation. It is however time to move to relational digital evidence, where recovered artefacts from different investigations can be cross referenced to provide new leads and establish patterns of activities of individuals and groups. This will allow the search and matching of evidence outwith a particular case and jurisdiction while observing privacy and ethical statutory requirements.

Keywords: very large scale forensics; big data;

I. INTRODUCTION

The original digital forensics discipline used to be simply “computer forensics” and until a decade ago it was the prevailing term for all digital forensic functions. With the advent of the smart phones we moved to mobile phone forensics, and currently the accepted convention is “digital forensics”, a term that encompasses the investigation of every different device and source of electronic evidence.

As the size and complexity of data increases, we see digital forensics tools (both commercial and open source) adapting to the way imaging and investigation of suspect devices is performed. The new approaches deal with the sheer volume and focus on minimizing the time scales required for the extraction, indexing and analysis of the recovered data.

Generally when indexing and analysis is completed, some of the data have been assigned evidential value, and are of use to the forensic reporting process. These findings are limited to the particular case and are only cross referenced with other devices related to the particular case or the individuals involved.

There are several models [1] of digital forensics investigations (Casey, 2011) but all describe the process from the incident and crime scene to the final reporting, for individual crimes. No evidence of work on cross jurisdiction evidence access and evaluation has been identified.

In a closed door session with cybercrime officers from police forces from Switzerland, Romania, France, Italy and Hungary, (Cyber Security Romania, Sibiu 2016) it was argued that the need of cross referencing actual evidence can be more important than simply adding people in a “most wanted” list. [2]

This position paper examines the premise that a new approach is needed for evidence extracted from computers. The proposed process of Very Large Scale Digital Forensics (VLSDF) is not based on the need to improve indexing and recovery, but on the fact that digital evidence needs to shift from linear investigations to the analysis of relational digital evidence, from multiple sources. The issues of jurisdiction, privacy and possible solutions are addressed.

II. BACKGROUND AND EXISTING TECHNOLOGY

A. Existing Technology

It is interesting that relevant technology already exists for a segment of the digital forensics discipline. This technology covers the investigation of mobile devices (mobile phones, tablets, satnavs).

A very good example of a tool that collates and visualises evidential data is XAMN [3] a proprietary
tool created by MSAB [4] and it is an add-on tool to their investigating platform XRY. [5]

Once at least two handsets have been investigated, the XAMN platform cross tabulates data that has been extracted and saved in a standardised format, irrespective of the handset’s platform (hardware and software). This particular visualisation model has proven to be useful when a large number of devices are investigated. So if a police force has examined a set of devices from different crime scenes, they can cross tabulate the recovered data to establish connections between apparently unrelated devices. These connections can be direct or indirect; in the example below, Device A has called or texted Device B. Device C has called or texted both Devices A and B (direct connections). Devices A, B and C have not contacted or been contacted by device D. However, Devices B and C have an indirect connection with Device D; the details of Contact X (indirect connection).

Fig. 1. Direct and indirect relationships between devices

Lee and Un [6] make the case for fast indexing of in a forensic search. They examined cloud computing and the use of Apache Hadoop treating digital evidence as Big Data. They proceeded in setting their own setup of a server and slaves using Hadoop, in order to outperform previous indexing benchmarks. This is the only reference to digital evidence used as Big Data so far. It looks at the problem of volume and processing time but is not dealing with cross-case evidence.

B. Computer forensics tools & limitations

Established digital forensic tools (i.e. EnCase [7], FTK [8], X-Ways [9]) already perform certain levels of indexing of recovered data. However this indexing has certain limitations:

The search between cases is possible only when cases are indexed by the same tool and the search is performed using actual search strings (i.e. an actual email). These limitations mean that existing digital evidence cannot be shared across jurisdictions as issues of privacy are involved.

A recent study of HM Chief Inspector of Constabulary [10] identified that many UK police forces had problems accessing specialist forensics services, and that delays in retrieving and analysing digital forensic evidence lead to unacceptable delays in prosecutions. The same report claims that digital forensics should be standardised and used the same way DNA and fingerprints are. This report supports the claim of this paper, that a large scale approach is now necessary.

III. MAKING THE CASE FOR VLSDF

The reason for the need of a VLSDF platform is that data needs to be cross tabulated and cross referenced with other data sets. This makes sense and it is feasible when it comes to mobile phones. The mobile phone forensic technology differs of course from the traditional computer forensics. It has to a certain degree a set of more standardised data sets of evidential value (i.e. SMS, call logs, GPS coordinates etc).

In order to make a case for the potential of a VLSDF platform, one could consider the following three hypothetical cases.

A. Case 1: Indecent images of children

An indecent image of a child is found in 4 different computers in three different jurisdictions. These jurisdictions do not share recovered evidence so no comparison can be performed between these pieces of evidence.

a. If these were a match, the following possible lines of investigation could have been pursued:

b. One of the computers was the source of the dissemination of the image.
c. Some or all of the users of these computers may have accessed the same website or downloaded the image from the same P2P source
d. One of the computers may be connected with the initial creation of the image
I. Case 2: An attached device

The serial number of an USB storage stick is recovered in five different computer examinations around Europe. These instances are in five different jurisdictions. This information is crucial to identify the individual who has had access to all five computers. The date stamps will create a timeline of the individual’s whereabouts.

II. Case 3: Undiscovered relationships

A computer is examined and extreme terrorist photographic material is recovered in the United States. The images contain metadata that include the serial number of the camera. A second computer is recovered in France. The computer is investigated for an unrelated offence (i.e. indecent images of children). A camera is seized together with the computer, or there is evidence of the camera being mounted on the computer using a USB data cable. The images recovered in France include government buildings in London. The serial number of the camera is recorded and a connection between the two can be established. The owner of the camera can be linked with the terrorist material in the US even if nothing is found on their camera or computer. The presence of images from London can trigger further investigation for potential attacks in the UK.

IV. PRIVACY AND LEGAL CONCERNS

Hu, Yu et al [11], make the case for the privacy of data found on data storage devices; especially the instances where the devices concerned are part of a server or a network, that many legitimate users access and whose privacy may be affected. The issue of privacy is important and different jurisdictions will be reluctant to share data and evidence between them.

Sharing simple sets of data like names, accounts etc is relatively easy, but allowing access to imaged data that may include sensitive information is practically impossible under the current frameworks. For example, if a private digital forensics consultant receives an imaged drive (from an English police force) as part of the legal process in England, then that consultant breaks no law in that jurisdiction. If that person with the imaged drive crosses the borders of Scotland, then they are in possession of indecent images of children and technically must be arrested. The Scottish judicial and police framework sees giving an imaged drive to an independent examiner outside the police or the court as “creating and distributing indecent images of children” and they simply refuse to do so. If there are gaps like this within the same country, ultra state sharing will be simply impossible. The privacy concerns and the legal issues between jurisdictions are the main counter arguments to sharing digital evidence.

V. IMPLEMENTATION

In order to navigate around legal restrictions, political decisions and jurisdictional differences, I propose a platform that will share anonymized information in the form of metadata.

Although one can see law enforcement as a major user of such platform, this can be implemented for large scale investigations by private consultants, and by international auditing firms who may be able to correlate evidence from different devices, different businesses and different countries.

A VLSDF platform should first of all conform to several standards. There will be a need for standardised extraction of metadata, categorization and storage of data types and other recovered information. Together with the metadata there will be a need for other data to be able to be categorised, and that will be a bigger challenge.

The initial schema for this platform should include:

i. A unique reference number for each source of data that will work like an EAN (International Article Number; barcode) or a VIN (Vehicle Identification Number).
ii. Definitions of metadata from images and videos that should be included in the evidence.
iii. An additional identifier of the source of an image or video (i.e. file downloaded from Facebook or Instagram).
iv. Definition of GPS metadata in various files
v. Definition of file types (technical) from file headers and other contact (i.e. JPEG header, quality factor of image etc.)
vi. Definition of text fields that will contain recovered text (natural language) from documents, emails, cached web pages etc.
vii. Serial numbers of devices attached to the computer (flash drives, mobile phones etc).
viii. IMEI (International Mobile Station Equipment Identity) numbers of mobile devices associated with the suspect computer
x. Internet browsing and FTP activity
xi. A checklist of installed and used applications that can be used for filtering later on. In this list if TOR is checked, then it will be easier to filter all cases where a TOR browser is installed, regardless of the amount of artefacts found.

xii. Some (or most) of the data sets can be hashed (i.e. a serial number, an email address, a user name) to achieve a greater degree of privacy.

Once the database is accessed by different users, they may search using different criteria including:

a. Keyword search (set by user)

b. Search of hash values (automated batch job); check prepared results

c. Search cases that have particular apps or other software installed

d. Search by coordinates and display any relevant hits on a proximity map. A reverse map visualisation tool may be developed, where the user selects a geographic area on a map and every piece of evidence from that area will be listed or pinned on the map. Selection limitations in place may reduce the risk of abusing the tool.

VI. LIMITATIONS

Like every application, the VLSDF platform has some limitations that one can foresee and most likely many that will appear in due course.

One of the major limitations is that the different set of data will probably need a different database, but this is something that can be dealt with.

The major drawback will be the way to insert digital artefacts in the database. Some can be fully automated, like metadata from files, image headers etc. Some others however will need human intervention. Free text is likely to be one of the biggest challenges. Personalised folder names and other data will need to be listed and selected by a user after some filters are applied. However some manual work (data entry) seems at this stage inevitable.

VII. CONCLUSIONS

It is this author’s belief that a VLSDF platform will be an invaluable tool in the search of cyber crime across investigations and across borders. It will also enhance the investigative capacity of other trans-border crimes, including (but not limited to) terrorism, human trafficking, illegal substances and weapons trafficking, fraud etc.

The implementation will have to be the result of long research both in academic and business environments. However at this early stage, I envisage two potential outcomes: The creation of a standardised database that is indexed and allows for large scale cross tabulations. The second possible outcome is the creation of a standardised format that will contain anonymized information in a way that is sufficient and easy to access, that could be implemented in existing applications or be converted to by a relevant tool.

REFERENCES