

Original citation:

Lallie, Harjinder Singh. (2014) The problems and challenges of managing crowd sourced audio-visual evidence. Future Internet, Volume 6 (Number 2). pp. 190-202.

Permanent WRAP url:

<http://wrap.warwick.ac.uk/60506>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions.

This article is made available under the Creative Commons Attribution- 3.0 Unported (CC BY 3.0) license and may be reused according to the conditions of the license. For more details see <http://creativecommons.org/licenses/by/3.0/>

A note on versions:

The version presented in WRAP is the published version, or, version of record, and may be cited as it appears here.

For more information, please contact the WRAP Team at: publications@warwick.ac.uk

warwick**publications**wrap

highlight your research

<http://wrap.warwick.ac.uk/>

Article

The Problems and Challenges of Managing Crowd Sourced Audio-Visual Evidence

Harjinder Singh Lallie

WMG (Warwick Manufacturing Group), University of Warwick, University Road, Coventry CV4 7AL, UK; E-Mail: h.s.lallie@warwick.ac.uk; Tel./Fax: +44-24-7655-1687

Received: 8 January 2014; in revised form: 25 February 2014 / Accepted: 26 February 2014 /

Published: 1 April 2014

Abstract: A number of recent incidents, such as the Stanley Cup Riots, the uprisings in the Middle East and the London riots have demonstrated the value of crowd sourced audio-visual evidence wherein citizens submit audio-visual footage captured on mobile phones and other devices to aid governmental institutions, responder agencies and law enforcement authorities to confirm the authenticity of incidents and, in the case of criminal activity, to identify perpetrators. The use of such evidence can present a significant logistical challenge to investigators, particularly because of the potential size of data gathered through such mechanisms and the added problems of time-lining disparate sources of evidence and, subsequently, investigating the incident(s). In this paper we explore this problem and, in particular, outline the pressure points for an investigator. We identify and explore a number of particular problems related to the secure receipt of the evidence, imaging, tagging and then time-lining the evidence, and the problem of identifying duplicate and near duplicate items of audio-visual evidence.

Keywords: time-lining; digital forensics; triage strategies; near-duplicate evidence

1. Introduction

The term “big data” generally refers to the problems of processing very large datasets often collected to finite detail and which can only be processed effectively through elaborate and sometimes complex techniques. The analysis and processing of such large data sets is problematic and is explored in a branch of data analysis often referred to as data analytics or data science.

The domain of data analytics advocates numerous data processing techniques which hitherto have not received the attention and focus in digital forensics that they require. That said, however, it is worth exploring whether a big data problem in digital forensics exists. The answer to the question is probably no, or at least we do not have enough evidence to suggest that there is a big data problem. However, we advocate that there is what we deem a “large data” problem which we explore in this study within the context of *crowd sourced audio-visual evidence*. This problem refers specifically to the complexities of receiving what could potentially be large numbers of data items, some of which may be duplicate or near-duplicate

2. What Is Big Data

To enable us to assert our view regarding the “large data” problem that we allude to within this paper, it helps to understand what big data is and, thereby, understand why the digital forensics community is most probably not facing a big data problem. What makes big data “big” is “repeated observations over time and/or space” and whose size “forces us to look beyond the tried-and-true methods that are prevalent at that time” [1]. Lynch [2] advocates that data can be “big” for a number of reasons for instance it: challenges the current boundaries of computational power and knowledge; is of lasting significance (the data will be being reused sometimes for many different reasons for a long time) and may present descriptive challenges in experimental set-up. The likelihood of reuse in different contexts raises design issues relating to the structure and shape of data being collected.

2.1. Volume, Velocity and Variety

McKinsey adds that the definition is contextually contemporaneous and relative, in other words what constitutes big data will change as the technology advances, “big data in many sectors today will range from a few dozen terabytes to multiple petabytes” [3]. Jacobs recollects that in the 1980s, the 100 GB hard disk enclosed in the IBM 3850 MSS (Mass Storage System) was used to provide researchers with ready access to the entire 1980 U.S. Census database, at the time it was considered to be a big data problem—although the term might not have been used in that context at the time [1].

The big data problem therefore comprises of three dimensions:

- **Volume** The size of data is too large (either or both in terms of number of items or size) to be processed effectively and efficiently.
- **Velocity** It takes too long to extract meaningful data from the dataset. This is a feature of volume and variety but additionally refers to unstructured data.
- **Variety** The dataset comprises of numerous complex structures of data and includes for instance: computer access logs, imagery, financial transactions, and website navigation trees.

For a big data problem to exist, it needs to present a problem across any of the three dimensions outlined above. In other words, data could be significant in size, but comprise of a single format which is easily understood and processed, in such a case we would not refer to it as big data.

Quite often, the problems of velocity and variety are a function of volume. Enterprises are collecting increasingly voluminous sizes of data. This is exemplified by a number of well-known cases. HP’s Neoview data warehouse processes around 267 million transactions generated daily by

4000 US Wal-Mart stores. This is in addition to the four petabytes of customer transaction data held by Wal-Mart [4]. The Large Hadron Collider produces around 15 petabytes of data per year. The data is accessed by thousands of scientists around the world and is stored in eleven “tier 1” computer centres each of which stores a large fraction of the data. The data is then made available to 160 “tier-2” centres for analysis [5]. CardioDX analyses clinical data from thousands of patients to build diagnostic algorithms used by physicians to determine the likelihood that patients had obstructive coronary artery disease [6].

2.2. Digital Investigative Trends

Whilst there appears little evidence in the digital forensics community concerning similarly large datasets, there do appear to be examples of very large and complex investigations.

The Internet Crime Complaint Center (IC3) receives and processes internet crime reports for the USA. It has reported a growth in the number of internet crimes being reported between 2007 and 2012 as follows: 2007: 206,884; 2008: 275,284; 2009: 336,665; 2010: 303,809; 2011: 314,246; 2012: 289,874 [7]. This does not necessarily represent a dramatic increase in internet crime reporting, in fact there was a slight drop in 2010 and 2012, the general trend however seems to be an increase. The two drops in reporting can be explained by the multiple agencies to whom crime can now be reported and by users being more aware of the range of agencies available for the filing of such complaints.

Similarly, the FBI has recorded a year on year growth in the number of digital investigations being conducted by them as well as a corresponding rise in the size of data being processed (Table 1 [8]).

Table 1. Number of e-Investigations undertaken by the Federal Bureau of Investigation (FBI) [8].

Year	Number of e-investigations	Size of data processed (TB)
2003	987	82.3
2004	1304	229
2005	2977	457
2006	3633	916
2007	4634	1288
2008	4524	1756
2009	6016	2334
2010	6564	3086
2011	7629	4263

The multitude of investigations by the FBI and other law enforcement authorities represent a problem across the dimensions of volume, velocity and variety. Social media site (SMS) investigations involve data that spans large networks and includes a multitude of objects such as links, actions, pictures, video, text and associations.

This is also demonstrated in large scale network investigations involving the analysis of corporate network log files such as those associated with firewall, IDS (Intrusion Detection Systems) or web servers. A typical organisational subdomain with 150 IP addresses may generate 60–70,000 IP entries in the firewall log per hour. Extend this to the whole network over the time period of a week and then across the range of log data available and this can easily reach more than 150 million entries.

The increasing complexity of digital investigative problems is evidenced further in a number of well-known case studies. Hans and Swinehart [9] highlight the example of the 2004 UN Oil for-food program fraud investigation. Iraq sold \$64.2 billion worth of oil to 278 companies, \$34.5 billion of the money was used to purchase humanitarian aid and goods from 3416 companies. The program was shrouded by allegations of bribes and fraud. The subsequent investigation involved thousands of documents and a substantial (but unstated) digital investigation. In this case, the financial scale of the fraud should not suggest that there was a correspondingly large and complex investigation, there may very well have been, however we do not have access to details relating to the investigation (as is often the case in such instances).

Possibly the most well-known large scale digital investigation is that of Enron. Following the collapse of the company in 2001–2002, the US Securities and Exchange Commission (SEC) and the Federal Energy Regulatory Commission (FERC) conducted two separate independent enquiries into the scale of the fraud undertaken by the company. The email mailboxes of more than 150 executives, comprising more than 600,000 emails, were investigated. At the time this would have been a very complex investigation, particularly with 600,000 emails involved and thousands of attachments and associated e-documents.

In 2012, the Computer Analysis Response Team, (CART—a department that provides assistance to the FBI in the search and seizure of digital evidence) supported around 14,000 investigations, conducted more than 133,000 digital investigations and analysed more than 10,500 Terabytes of data [10]. However, whilst this data size appears large, it is related to 14,000 investigations amounting to around 0.7 TB (700 GB) per investigation—which in the present day and age is not necessarily voluminous. In the same year, the FBI was ordered to copy 150 terabytes of data held on the MegaUploads server by Kim Dotcom [11]. In 2013, Ian Watkins was found guilty of possessing up to 27 TB of paedophile images in what was one of the largest hauls by South Wales police in the UK [12]. Ian had used online cloud storage and the collection amounted to more data than the police authority held in its own data storage systems.

In both the cases of Kim Dotcom and Ian Watkins, we should note that whilst the data size appears large, these do not appear to be a “needle in the haystack” situation. In the case of Kim Dotcom, prosecutors would reasonably easily have confirmed the existence of films/audio and quite easily (it would seem) been able to confirm the transmission of these files. In the case of Ian Watkins, the defendant pleaded guilty when the weight of evidence was presented to him, thereby avoiding a trial. If the case had gone to trial, investigators may not have needed to plough through the 27 TB of data to find sufficient evidence to prosecute. The difficulty in both these cases and increasingly so in many more cases is the challenge of imaging the increasingly large datasets in a timely manner.

The increasing complexity of investigations is also highlighted by the growing number of cloud investigations taking place. Cloud investigation has not received the research interest that it deserves, Beebe [13] and ENISA (European Network and Information Security Agency) [14] highlighted the need to prioritise further research into cloud investigation and in particular evidence gathering mechanisms. Grispos *et al.* concluded that current methods and guidelines for digital investigation could be insufficient for conducting a cloud investigation [15].

Possibly one of the biggest problems in investigating the cloud is that of identifying and then subsequently imaging potentially large data sources. A public cloud storage infrastructure may consist

of dozens of server farms/data stores located at different geographic locations against which the data may be dynamically routed and stored [16]. The investigator has to identify the precise location of the data before being able to image the data. Time-lining is quite fundamental to a digital investigation; however the uncertainties surrounding the location of data make it more difficult to timeline. File metadata does not store information relating to its movement and an investigator may struggle to chart the movement of data over any given period.

3. “Large Data” and Digital Forensics

The above digital investigative examples are complex investigative problems, but are not big data problems and the examples given therein are quite manageable, for instance, despite the large number of entries in an organisational log file, the digital investigative industry possesses the tools and techniques to be able to identify patterns, trends and problems within such log files. Furthermore, the increase in data being investigated by law enforcement agencies as reported by the FBI can be dealt with using techniques such as digital forensic triage—assuming the law enforcement agency concerned is prepared to accept triage as a viable technique for prioritising investigation workload.

It is in fact difficult to ascertain whether there is a big data problem in digital forensics particularly because little if any data is publicly available regarding the size, nature and complexity of law enforcement investigations. It is very rare for a digital investigation agency to publish details regarding the challenges and problems faced during an individual investigation or about the size of data recovered and subsequently investigated. Should such data be available we would be able to better determine whether or not a big data problem exists.

However, we propose within this paper that many investigations have the capacity to be large enough to be considered—as we term it, “large data” problems, these are investigative problems which can take years to conclude. One particular example of such large investigations involves the processing of *crowd sourced audio-visual evidence* which creates major technical and jurisprudence challenges.

4. Crowd Sourced Audio-Visual Evidence

Crowd sourced audio-visual evidence is a phenomenon where law enforcement agencies collect video, audio and photographic footage (collectively referred to herein as audio-visual) from the public to aid in the investigation of tempestuous events such as civil unrest and riots. Whilst the use of crowd sourced audio-visual evidence has not yet been expanded beyond this, it could also serve as a useful source of evidence in many other incidents wherein footage taken by the public on or around the time of an incident could prove useful in a subsequent investigation and can augment evidence already gathered through traditional systems such as CCTV.

Crowd sourced audio-visual evidence has a value outside the domain of law enforcement, for instance, it is of great value to news agencies who have relied on crowd sourced audio-visual evidence for years. Such evidence was used during a number of recent international events such as the conflicts, civil disorders and uprisings in Egypt, Syria and Turkey. However, like the law enforcement agencies, the news agencies must validate and authenticate the submitted evidence and in the absence of such authentication, they present it with the disclaimer/caveat that the evidence is “unverified footage” or is based on “unconfirmed reports”.

This form of footage is invaluable to governments and responder agencies such as the United Nations who often gain insights into world events through such reporting. However, they are unable to react to serious incidents until such footage and particularly the incidents the footage purports to represent are confirmed and validated.

There have been two prominent recent examples of the use of crowd sourced audio-visual evidence by law enforcement agencies: the Victoria Stanley Cup riots in Vancouver (2011) and the UK riots of the same year.

4.1. Victoria Stanley Cup Riots 2011

In 2011, the Vancouver Police Department (VPD) encouraged the public to post to their Twitter feed whilst the VPD were managing the Stanley Cup playoffs. The Twitter feed allowed the public to engage with the authorities whilst at the same time allowing them to monitor unrest. In June of the same year—by which time VPD had more than 16,000 Twitter followers, a riot broke out following a game. Videos of the riots were captured by attendees to the event as well as passers-by on their mobile phones.

Witnesses began to post videos of the riot to websites, blogs, Twitter feeds, and SMSs and proceeded to “tag” and thereby identify potential perpetrators. Further to this, witnesses posted more than 1000 emails to the VPD, this was in addition to the 10,000+ tweets posted during and after the event. The VPD recognized the evidentiary value of this crowd sourced evidence and encouraged the public to send these in to form part of the evidence being considered in the investigation.

It took the VPD some weeks to establish a process which allowed citizens to send in the videos. The size of potential evidence submitted was staggering and by VPD’s admission, it would take two years to process the data [17]. In fact, two years later, rioters were still being prosecuted following the investigation of individual cases.

By 20 July, the VPD had acquired 1,500 hours of video, 15,000 images/photographs and more than 4,000 email messages with potential evidence. During the intermitting time period, a server hosting VPD press releases crashed because of the weight of traffic being generated from the on-going Twitter feeds relating to the event.

This case study forms one of the first examples of the value and use of crowd sourced audio-visual evidence. It represents a problem of volume and velocity and less so of variety. The major form of evidence were videos, however there were a lot of them and the time frame within which to prosecute put pressure on the investigating authorities to show that justice was being done.

4.2. London Riots

Months later in August 2011, the UK was subjected to the biggest riots the country had seen in decades. During the riots centred around London, Manchester, Liverpool, Birmingham, and Bristol, more than 150 police officers were injured and five people died. Three thousand, four hundred and forty three riot related crimes were reported in London alone resulting in more than £200 million of damage caused in the capital. Three thousand, one hundred people were arrested. In this case, SMSs were used at both ends of the law by rioters who organised themselves through SMS and by citizens—who as in the Stanley Cup riots, posted videos and messages after they witnessed incidents and to

report incidents. For a little while, the Government considered banning people from using SMSs if they were suspected to be involved in plotting criminal activity [18].

The UK has over 1.85 million CCTV cameras in operation (over 130,000 of which capture 29 megapixel images) [19,20]. The Metropolitan Police Service (MPS) estimated that there were 200,000 hours of CCTV footage linked to the riot and by March 2012 (seven months later) only 75% of footage had been viewed. Five thousand image evidence packages had been produced, 1,100 of which have had the offenders fully confirmed and identified [21].

To cope with the size of the investigation, the MPS purchased an additional 149 CCTV viewing stations and trained 83 police officers and 100 volunteers to undertake the CCTV investigations. A bespoke system designed to allow efficient cataloguing and searching of wanted images was purchased [21].

In addition to the evidence that existed in CCTV and other systems operated by the authorities, the public was encouraged to send in their own footage, resulting in 1000 s of hours of extra footage for processing.

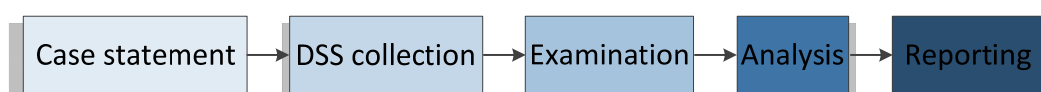
5. The Problems of Crowd Sourced Audio-Visual Evidence

These two cases demonstrate the value of crowd sourced audio-visual evidence—particularly in augmenting evidence collected by law enforcement agencies. However it also presents important challenges. The acquisition of crowd sourced audio-visual evidence can result in 1000s of hours of footage which has to be authenticated, time-lined and then analysed. For such a source of evidence to be considered useful, agencies must outline robust procedures for managing the evidence.

So what are the digital investigative problems relating to crowd sourced audio-visual evidence? To proceed, we must consider the digital investigation process as outlined in Figure 1. Whilst there is no universally accepted framework for investigation, a typical investigation begins with a “case statement” which details the nature of the case and in particular the intended outcome of the investigation. The investigation typically involves four phases defined by the Association of Chief Police Officers (ACPO) [22] and the National Institute of Justice (NIJ) [23] which are the: collection phase (the collection and documentation of potential evidence), examination phase (which seeks to identify important evidence), analysis phase (which seeks to establish the significance and probative value of evidence extracted) and the reporting phase (which highlights the key findings of the investigation).

Crowd sourced audio-visual evidence presents challenges in the collection, examination and analysis phases of an investigation specifically in terms of acquisition (of a potentially large number of datasets), the appropriate and proper tagging of each acquired image with details of the source of the evidence, handling near duplicate data and time-lining complex sequences of data.

Figure 1. Digital Investigation Process.



5.1. Collecting and Managing Crowd Sourced Audio-Visual Evidence

An important problem with crowd sourced evidence relates to the process of receiving, imaging and managing what may become significant numbers of evidence items from numerous sources.

In a normal investigation, this involves attending a “crime scene”, seizing, bagging/tagging and then imaging the evidence. In crowd sourced audio-visual evidence, the evidence sources are remotely located and just as evidence from a crime scene must be protected from the likelihood of being tainted or corrupted, so must evidence supplied remotely. This implies that the evidence must be received over a secure channel using protocols such as SSL/HTTPS. This does not of course prevent the tainting and corruption of evidence at either end of the transmission.

This process must not only be secure, but also be seen to be secure; in other words, users must have confidence in the security as otherwise anxious users may feel uncomfortable in submitting potentially incriminating evidence. This implies that the whole system of encouraging citizens to submit evidence through to the receipt and management of evidence must contain visual and verbal cues that instil confidence in those submitting evidence.

Added to this—and learning from the VPD experience, we have the problem of concurrency and load management/balancing. The receiving servers must be designed to withstand large and numerous concurrent data uploads in potentially sort periods of time.

5.2. Acquisition

Once the evidence is received, it must be imaged. Digital investigations are rarely if ever performed on the original digital storage systems (DSS) as these must be preserved to ensure evidence integrity. The investigator makes a digital forensic image (DFI) of the DSS and the investigation proceeds on the DFI.

By July 20th, 1500 hours of video footage of the Stanley Cup riots were received. This number most likely increased as the investigation progressed. This could amount to hundreds—possibly thousands of individual videos each of which must be imaged.

The problems associated with imaging large datasets has been long recognised [24] and whilst hard drive speeds and data capture speeds have increased (implying an increased rate of data acquisition), the corresponding increase in hard disk capacity means that DSS acquisition time has actually increased. So for instance previous capture speeds of 58 MB/s allowed for a 200 GB hard disk to be captured in one hour, whilst capture speeds have increased to—for instance 128 MB/s, the corresponding increase in hard disk storage means that it takes around seven hours to capture a 3 TB [25].

In the case of crowd sourced audio-visual evidence, the investigator may not know the full extent of data to be acquired and the longer an investigation proceeds, the greater the evidence pool may become.

5.3. Tagging

The process of tagging crowd sourced audio-visual evidence is not much different from evidence tagging and bagging in a typical digital investigation, however there are a number of particular nuances of which the investigator must be aware. For instance one must record: details of the sender

(for jurisprudence purposes—should the sender be required to give evidence) as well as technical details relating to the device on which the audio-visual was recorded such as name/model number, operating system, app that recorded the video and the date/time and the location the video was taken.

There are two sources for this data, it may be contained within the metadata attached to the submitted file—but that assumes that the device is capable to recording this data and that the user has enabled the device to record metadata. In addition to this, the user could be asked to supply the data and, in the event that both metadata and user-supplied data are available, the two be combined/augmented to present a more reliable tag for the evidence.

But here again we are presented with issues of usability. A submission screen laden with user prompts requesting technical data relating to the submitted audio-visual file is likely to discourage respondents.

5.4. The Problem of Near Duplicate Audio-Visual Evidence

The propensity of SMSs means that there are likely to be numerous versions of individual audio-visual evidence submitted. Users may have found videos of evidence in their SMS space and may feel compelled to submit it in response to a call for evidence. This results in duplicate or near duplicate versions of an audio-visual. Whilst hash checking will highlight duplicate data, it does not cater for near duplicate audio-visuals.

To understand the problem of near-duplicate evidence, consider this. Digital investigators rely on hashes—also referred to as message authentication codes (MAC) to identify identical data. A single bit change between two identical items of data results in a completely different hash, in such cases where two items of audio-visual are essentially duplicates but have had a small change made to them, we refer to these as *near duplicate*. Examples of this can include a colour photograph i_1 that has been converted to black and white/sepia or has had its format converted from BMP to TIFF to make i_2 , or a video v_1 that has had the audio quality reduced or its version changed from AVI to MKV to produce v_2 . In this case, i_1 and i_2 are near duplicate as are v_1 and v_2 .

In other cases, data may deliberately have been altered to avoid detection, for instance law enforcement agencies use hash databases to speed up the process of photograph grading in child exploitation cases. It is not uncommon for suspects to have deliberately altered single bits in images to avoid detection. This is also common in more simple efforts to avoid spam/malware detection filters.

A failure on the part of a digital investigator to recognise that two items of data are duplicates or near duplicates can have a considerable impact on an investigation; hence a lot of research has focused on the detection of near duplicate images [26,27], videos [28] and audio [29–31].

Near duplicate detection is most difficult for video, this is because of the complexity of video as compared to still images and audio. Videos footage can transgress multiple dimensions—for instance it comprises of thousands of photo frames (in some cases tens of thousands), it can contain multiple audio tracks, subtitles and other rich information. A change in any single bit within one of these elements makes the video a near duplicate.

A number of efforts have been deployed to address this and these include correlation based detection systems [32], the Hadoop MapReduce programming model [33] and technologies that look at the constituent elements of an image (from which the video for instance is constructed), PhotoDNA is

an example of the latter [34]. PhotoDNA identifies the most prominent contrast edges—called intensity gradients in each image and uses these to compare images with each other.

Time-Lining

Whilst the actual analysis of the acquired video footage will most likely be done by a case investigator or video forensics expert, the digital investigator may very well be involved with time-lining the acquired video footage. Time-lining audio-visual evidence acquired from multiple sources provides the investigator with a unique perspective of the evidence dissimilar to almost any investigations of which the investigator may be aware. The situation almost resembles having multiple “high-quality” CCTV angles.

Notwithstanding the “rich” evidence source available, time-lining thousands of hours of acquired video footage presents a considerable logistical challenge and this becomes more problematic where the time period to which the incident relates is quite short. The London riots lasted for days, whereas the Boston bombing lasted moments (although there was plenty of periphery evidence relating to the movements of the crowds in the run-up to the incident).

Hence the investigator may be presented with video footage from multiple angles, places and times, possibly covering a very short period of time. This footage has to be synchronised to the timeline to present an accurate temporal view of the sequence of events. One way to do this is to consider the time data contained in the metadata of the submitted file if it exists. A second, perhaps complimentary approach, is to outline benchmark footage for which the time can be authenticated and then to calibrate other footage against the benchmark data.

6. Conclusions

In this paper we have explored the benefits, challenges and problems associated with the use of crowd sourced audio-visual evidence in digital investigations. We have posited that whilst there is no clear evidence of a big data problem existing in digital forensics, we do have examples of highly complex investigations which we deem to be “large data” problems.

The use of crowd sourced audio-visual evidence is an example of a large data problem and is associated with a series of problems, such as the secure receipt of the evidence, its acquisition, the efficient tagging and subsequent time-lining of the evidence and the problem of identifying duplicate and near duplicate items of audio-visual evidence.

There are a number of areas in which this work and research can be further progressed. We are as yet unable to determine the extent of the digital investigation problem, particularly in terms of its size and complexity; it would be useful for instance to determine whether there is a “large data” problem. It would be useful to further explore strategies for time-lining audio-visual evidence particularly against its temporal dimension. Finally, the problem of near-duplicate evidence needs to be further explored against the dimensions of audio, visual and photographic evidence.

Conflicts of Interest

“The author declares no conflict of interest”.

References

1. Jacobs, A. The pathologies of big data. *Commun. ACM* **2009**, *52*, 36–44.
2. Lynch, C. Big data: How do your data grow? *Nature* **2008**, *455*, 28–29.
3. Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Byers, A.H. *Big Data: The Next Frontier for Innovation, Competition and Productivity*; Report from McKinsey Global Institute, May 2011. Available online: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed on 4 March 2014).
4. Weier, M.H. In A Big Win For HP, Wal-Mart Chooses Neoview Data Warehouse. 2007. Available online: <http://www.informationweek.com/news/201202317> (accessed on 2 June 2012).
5. European Organization for Nuclear Research. Worldwide LHC Computing Grid. 2008. Available online: <http://public.web.cern.ch/public/en/lhc/Computing-en.html> (accessed on 2 June 2012).
6. Revolution Analytics. Complex Data Sets in Genomic Diagnostics Require Multiple Analytic Methods. 2010. Available online: <http://www.revolutionanalytics.com/content/complex-data-sets-genomic-diagnostics-require-multiple-analytic-methods> (accessed on 2 June, 2012).
7. Internet Crime Complaint Center. *Internet Crime Report*; Internet Crime Complaint Center: Washington D.C., Washington, USA, 2012.
8. Regional Computer Forensics Laboratory (RCFL), Federal Bureau of Investigation (FBI), U.S. Department of Justice. *Annual Report for Fiscal Year 2011*; FBI: Washington, DC, USA, 2011.
9. Hans, S.; Swinehart, G. *Finding the Needle—Using Forensic Analytics to Understand What Happened—and What Might Happen*; Deloitte: Boston, Massachusetts, USA, 2011.
10. Federal Bureau of Investigation (FBI). Piecing Together Digital Evidence—The Computer Analysis Response Team. 2013. Available online: <http://www.fbi.gov/news/stories/2013/january/piecing-together-digital-evidence/piecing-together-digital-evidence> (accessed on 5 March 2014).
11. FBI ordered to copy seized Dotcom data. *Otago Daily Times*, 20 January 2012. Available online: <http://www.odt.co.nz/news/national/213394/fbi-ordered-copy-seized-dotcom-data> (accessed on 5 March 2014).
12. Lostprophets’ Ian Watkins: “Tech savvy” web haul. *BBC News*, 18 December 2013. Available online: <http://www.bbc.co.uk/news/uk-wales-25435751> (accessed on 1 March 2014).
13. Beebe, N. Digital forensic research: The good, the bad and the unaddressed. In *Advances in Digital Forensics V*, Proceedings of the Fifth IFIP WG 11.9 International Conference on Digital Forensics, Orlando, FL, USA, 26–28 January 2009; Springer Berlin Heidelberg: Berlin/Heidelberg, Germany, 2009; Volume V, pp. 17–36.
14. European Network and Information Security Agency (ENISA). Cloud Computing. Benefits, risks and recommendations for information security. 2009. Available online: http://www.enisa.europa.eu/activities/risk-management/files/deliverables/cloud-computing-risk-assessment/at_download/fullReport (accessed on 2 June, 2012).
15. Grispos, G.; Storer, T.; Glisson, W.B. Calm Before the Storm: The Challenges of Cloud Computing in Digital Forensics. *Int. J. Digit. Crime Forensics* **2012**, *4*, 28–48.
16. Qureshi, A. Plugging into Energy Market Diversity. In Proceedings of the 7th ACM Workshop on Hot Topics in Networks, Calgary, Canada, 6–7 October 2008.

17. *2011 Stanley Cup Riot Review*; Vancouver Police Department: Vancouver, Canada, 2011.
18. Halliday, J. David Cameron considers banning suspected rioters from social media. *The Guardian*, 11 August 2011.
19. High-def CCTV cameras risk backlash, warns UK watchdog. *BBC News*, 3 October 2010. Available online: <http://www.bbc.co.uk/news/technology-19812385> (accessed on 2 June, 2012).
20. Lewis, P. You're being watched: There's one CCTV camera for every 32 people in the UK. *The Guardian*, 2 March 2011.
21. *Metropolitan Police Service. 4 Days in August. Strategic Review into the Disorder of August 2011*; Metropolitan Police: London, UK, 14 March 2012.
22. *Good Practice Guide for Computer Based Electronic Evidence*; Association of Chief Police Officers: London, UK, 2007.
23. National Institute of Justice. *Electronic Crime Scene Investigation: A Guide for First Responders*, 1st ed.; National Institute of Justice, U.S. Department of Justice: Washington, DC, USA, 2001.
24. Richard, G.G., III; Roussev, V. Next-generation digital forensics. *Commun. ACM* **2006**, *49*, 76–80.
25. Roussev, V.; Quates, C.; Martell, R. Real-time digital forensics and triage. *Digit. Investig.* **2013**, *10*, 158–167.
26. Ke, Y.; Sukthankar, R.; Huston, L. Efficient Near-Duplicate Detection and Sub-Image Retrieval. In Proceedings of the 12th ACM International Conference on Multimedia (ACM Multimedia 2004), New York, NY, USA, 10–16 October 2004; p. 5.
27. Chum, O.; Philbin, J.; Zisserman, A. Near Duplicate Image Detection: min-Hash and tf-idf Weighting. In Proceedings of the 19th British Machine Vision Conference (BMVC 2008), Leeds, UK, 1–4 September 2008; pp. 812–815.
28. Stamm, M.C.; Lin, W.S.; Liu, K.R. Temporal forensics and anti-forensics for motion compensated video. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 1315–1329.
29. Cano, P.; Batle, E.; Kalker, T.; Haitsma, J. A Review of Algorithms for Audio Fingerprinting. In Proceedings of 2002 IEEE Workshop on Multimedia Signal Processing, St. Thomas, Virgin Islands, USA, 9–11 December 2002; pp. 169–173.
30. Haitsma, J.; Kalker, T. A Highly Robust Audio Fingerprinting System. In Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002), Paris, France, 13–17 October 2002; pp. 107–115.
31. Cano, P.; Batle, E.; Gómez, E.; de Campos Teixeira Gomes, L.; Bonnet, M. Audio fingerprinting: concepts and applications. In *Computational Intelligence for Modelling and Prediction*; Halgamuge, S.K., Wang, L., Eds.; Springer: Berlin, Germany, 2005; pp. 233–245.
32. Liu, J.; Huang, Z.; Shen, H.T.; Cui, B. Correlation-based retrieval for heavily changed near-duplicate videos. *ACM Trans. Inf. Syst.* **2011**, *29*, pp 21–46.
33. Wang, H.; Shen, Y.; Wang, L.; Zhufeng, K.; Wang, W.; Cheng, C. Large-scale multimedia data mining using MapReduce framework. In Proceedings of the IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom), Taipei, Taiwan, 3–6 December 2012; pp. 287–292.

34. PhotoDNA Newsroom. *Microsoft News Center*, 10 September 2013. Available online: <http://www.microsoft.com/en-us/news/presskits/photodna/> (accessed on 10 November 2013).

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).