Digital Forensics as a Big Data Challenge

Alessandro Guarino

StudioAG
a.guarino@studioag.eu

Abstract

Digital Forensics, as a science and part of the forensic sciences, is facing new challenges that may well render established models and practices obsolete. The dimensions of potential digital evidence supports has grown exponentially, be it hard disks in desktop and laptops or solid state memories in mobile devices like smartphones and tablets, even while latency times lag behind. Cloud services are now sources of potential evidence in a vast range of investigations and network traffic also follows a growing trend and in cyber security the necessity of sifting through vast amount of data quickly is now paramount. On a higher level investigations – and intelligence analysis – can profit from sophisticated analysis of such datasets as social network structures, corpora of text to be analysed for authorship and attribution. All of the above highlights the convergence between so-called data science and digital forensics, to tack the fundamental challenge of analyse vast amount of data ("big data") in actionable time while at the same time preserving forensic principles in order for the results to be presented in a court of law. The paper, after introducing digital forensics and data science, explores the challenges above and proceed to propose how techniques and algorithms used in big data analysis can be adapted to the unique context of digital forensics, ranging from the managing of evidence via Map-Reduce to machine learning techniques for triage and analysis of big forensic disk images and network traffic dumps. In the conclusion the paper proposes a model to integrate this new paradigm into established forensic standards and best practices and tries to foresee future trends.

1 Introduction

1.1 Digital Forensics

What is digital forensics? We report here one of the most useful definitions of digital forensics formulated. It was developed during the first Digital Forensics Research Workshop (DFRWS) in 2001 and it is still very much relevant today:

> *Digital Forensics is the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.[Pear01]*

This formulation stresses first and foremost the scientific nature of digital forensics methods, in a point in time when the discipline was transitioning from being a "craft" to an established field and rightful part of the forensic sciences. At that point digital forensics was also transitioning from being mainly practised in separated environments such as law enforcement bodies and enterprise audit offices to a unified field. Nowadays this process is very advanced and it can be

H. Reimer et al. (Eds.), *ISSE 2013 Securing Electronic Business Processes,*
DOI 10.1007/978-3-658-03371-2_17, © Springer Fachmedien Wiesbaden 2013

said that digital forensics principles, procedures and methods are shared by a large part of its practitioners, coming from different backgrounds (criminal prosecution, defence consultants, corporate investigators and compliance officers). Applying scientifically valid methods implies important concepts and principles to be respected when dealing with digital evidence. Among others we can cite:

- Previous validation of tools and procedures. Tools and procedures should be validated by experiment prior to their application on actual evidence.
- Reliability. Processes should yield consistent results and tools should present consistent behaviour over time.
- Repeatability. Processes should generate the same results when applied to the same test environment.
- Documentation. Forensic activities should be well-documented, from the inception to the end of evidence life-cycle. On one hand strict chain-of-custody procedures should be enforced to assure evidence integrity and the other hand complete documentation of every activity is necessary to ensure repeatability by other analysts.
- Preservation of evidence – Digital evidence is easily altered and its integrity must be preserved at all times, from the very first stages of operations, to avoid spoliation and degradation. Both technical (e.g. hashing) and organizational (e.g. clear accountability for operators) measures are to be taken.

These basic tenets are currently being challenged in many ways by the shifting technological and legal landscape practitioners have to confront with. While this paper shall not dwell much on the legal side of things, this is also obviously something that is always to be considered in forensics.

Regarding the phases that usually make up the forensic workflow, we refer here again to the only international standard available[ISO12] and describe them as follows:

- Identification. This process includes the search, recognition and documentation of the physical devices on the scene potentially containing digital evidence.[ISO12]
- Collection – Devices identified in the previous phase can be collected and transferred to an analysis facility or acquired (next step) on site.
- Acquisition – This process involves producing an image of a source of potential evidence, ideally identical to the original.
- Preservation – Evidence integrity, both physical and logical, must be ensured at all times.
- Analysis – Interpretation of the data from the evidence acquired. It usually depends on the context, the aims or the focus of the investigation and can range from malware analysis to image forensics, database forensics, and a lot more of application-specifics areas. On a higher level analysis could include content analysis via for instance forensics linguistics or sentiment analysis techniques.
- Reporting – Communication and/or dissemination of the results of the digital investigation to the parties concerned.

1.2 Data Science

Data Science is an emerging field basically growing at the intersection between statistical techniques and machine learning, completing this toolbox with domain specific knowledge, having as fuel big datasets. Hal Varian gave a concise definition of the field:

[Data science is] the ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it.[Vari09]

We can see here the complete cycle of data management and understand that data science in general is concerned with the collection, preparation, analysis, visualization, communication and preservation of large sets of information; this is a paraphrase of another insightful definition by Jeffrey Stanton of Syracuse University School of Information Studies. The parallels with the digital forensics workflow are clear but the mention in both definition of visualization deserves to be stressed. Visualization is mostly never mentioned in digital forensics guidelines and standards but as the object of analysis move towards "Big Data", it will necessarily become one of the most useful tools in the analyst's box, for instance in the prioritization phase but also for dissemination and reporting: visual communication is probably the most efficient way into a human's brain but this channel is underused by most of today forensic practitioners.

If Data Science is concerned with "Big Data", what is Big Data anyway? After all big is a relative concept and prone to change with time. Any data that is difficult to manage and work with, or in other words datasets so big that for them conventional tools – e.g. relational databases – are not practical or useful.[ISAC13] From the point of view of data science the challenges of managing big data can be summarized as three Vs: Volume (size), Velocity (needed for interactivity), Variety (different sources of data). In the next paragraph we shall see how this three challenges dovetail nicely with the digital forensics context.

2 Challenges

"Golden Age" is a common definition for the period in the history of digital forensics that went roughly from the 1990s to the first decade of the twenty-first century. During that period the technological landscape was dominated by the personal computer, and mostly by a single architecture – x86 plus Windows – and data stored in hard drives represented the vast majority of evidence, so much so that "Computer Forensics" was the accepted term for the discipline. Also the storage size allowed for complete bitwise forensic copies of the evidence for subsequent analysis in the lab. The relative uniformity of the evidence nature facilitated the development of the digital forensic principles outlined above and enshrined in several guidelines and eventually in the ISO/IEC 27037 standard. Inevitably anyway they lagged behind the real-world developments: recent years brought many challenges to the "standard model", first among them the explosion in the average size of the evidence examined for a single case. Historical motivations for this include:

- A dramatic drop of hard drives and solid state storage cost (currently estimated at $80 per Terabyte) and consequently an increase in storage size per computer or device;
- Substantial increase in magnetic storage density and diffusion of solid-state removable media (USB sticks, SD and others memory cards etc) in smartphones, notebooks, cameras and many other kinds of devices;
- Worldwide huge penetration of personal mobile devices like smartphones and tablets, not only in Europe and America, but also in Africa – where they constitute the main communication mode in many areas – and obviously in Asia;
- Introduction and increasing adoption by individuals and businesses of cloud services – Infrastructure services (IAAS), platform services (PAAS) and applications (SAAS) – made possible in part by virtualization technology enabled in turn by the modern multi-core processors;

- Network traffic is ever more part of the evidence in cases and the sheer size of it has – again – obviously increased in the last decade, both on the Internet and on 3G-4G mobile networks, with practical but also ethical and political implications;
- Connectivity is rapidly becoming ubiquitous and the "Internet of things" is near, especially considering the transition to IPv6 in the near future. Even when not networked, sensors are everywhere, from appliances to security cameras, from GPS receivers to embedded systems in cars, from smart meters to Industrial Control Systems.

To give a few quantitative examples of the trend, in 2008 the FBI Regional Computer Forensics Laboratories (RCFLs) Annual Report[FBI08] explained that the agency's RCFLs processed 27 percent more data than they did during the preceding year; the 2010 Report gave an average case size of 0.4 Terabytes. According to a recent (2013) informal survey among forensic professionals on Forensic Focus, half of the cases involve more than on Terabyte of data, with one in five over five Terabytes in size.

The simple quantity of evidence associated to a case is not the only measure of its complexity and the growing in size is not the only challenge that digital forensics is facing: evidence is becoming more and more heterogeneous in nature and provenience, following the evolving trends in computing. The workflow phase impacted by this new aspect is clearly analysis where, even when proper prioritization is applied, it is necessary to sort through diverse categories and source of evidence, structured and unstructured. Data sources themselves are much more differentiated than in the past: it is common now for a case to include evidence originating from personal computers, servers, cloud services, phones and other mobile devices, digital cameras, even embedded systems and industrial control systems. File formats

3 Rethinking Digital Forensics

In order to face the many challenges but also to leverage the opportunities it is encountering the discipline of digital forensics have to rethink in some ways established principles and reorganize well-known workflows, even include and use tools not previously considered viable for forensic use – concerns regarding the security of some machine learning algorithms has been voiced, for instance in [BBC+08]. On the other hand forensic analysts' skills need to be rounded up to make better use of these new tools in the first place but also to help integrate them in forensic best practices and validate them. The dissemination of "big data" skills will have to include all actors in the evidence lifecycle, starting with Digital Evidence First Responders (DEFRs), as identification and prioritization will see their importance increased and skilled operators will be needed from the very first steps of the investigation.

3.1 Principles

Well-established principles shall need to undergo at least a partial extension and rethinking because of the challenges of Big Data.

- Validation and reliability of tools and methods gain even more relevance in a big data scenario because of the size and variety of datasets, coupled with the use of cutting-edge algorithms that still need validation efforts, including a body of test work first on methods and then on tools in controlled environments and on test datasets before their use in court.

- Repeatability has long been a basic tenet in digital forensics but most probably we will be forced to abandon it, at least in its strictest sense, for a significant part of evidence acquisition and analysis. Already repeatability stricto sensu is impossible to achieve in nearly all instance of forensic acquisition of mobile devices and the same applies to cloud forensics. When Machine Learning tools and methods become widespread, reliance on previous validation will be paramount. As an aside, this stresses once more the importance of using open methods and tools that can be independently and scientifically validated as opposed to black box tools or – worse – LE-reserved ones.
- As for documentation, its importance for a sound investigation is even greater when we see non-repeatable operations and live analysis routinely be part of the investigation process. Published data about validation results of tools and methods used – or at least pointers to it – should be integral part of the investigation report.

3.2 Workflow

Keeping in mind how the forensic principles may need to evolve, we present here a brief summary of the forensics workflow and how each phase may have to adapt to big data scenarios. ISO/IEC 27037 International Standard covers the identification, collection, acquisition and preservation of digital evidence (or, literally, "potential" evidence). Analysis and disposal are not covered by this standard, but will be in future – in development – guidelines in the 27xxx series.

Identification and collection
Here the challenge is selecting evidence timely, right on the scene. Guidelines for proper prioritization of evidence should be further developed, abandoning the copy-all paradygm and strict evidence integrity in favor of appropriate triage procedures: this implies skimming through all the (potential) evidence right at the beginning and selecting relevent parts. First responders' skills will be even more critical that they currently are and, in corporate environments, also preparation procedures.

Acquisition
When classic bitwise imaging is not feasible due to the evidence size, prioritization procedures or "triage" can be conducted, properly justified and documented because integrity is not absolute anymore and the original source has been modified, if only by selecting what to acquire. Visualization can be a very useful tool, both for low-level filesystem analysis and higher level content analysis. Volume of evidence is a challenge because dedicated hardware is required for acquisition – be it storage or online traffic – while in the not so distant past an acquisition machine could be built with off-the-shelf hardware and software. Variety poses a challenge of a slightly different kind, especially when acquiring mobile devices, due to the huge number of physical connectors and platforms.

Preservation
Again, preservation of all evidence in a secure way and complying with legal requirements, calls for quite a substantial investment for forensic labs working on a significant number of cases.

Analysis
Integrating methods and tools from data science implies surpassing the "sausage factory" forensics still widespread today, where under-skilled operators rely heavily on point and click all-in-one tools to perform the analysis. Analysts shall need to include a plurality of tools in their

panoply and not only that, but understand and evaluate the algorithms and implementations they are based upon. The absolute need for highly skilled analysts and operators is clear, and suitable professional qualifications will develop to certify this.

Reporting
The final report for an analysis conducted using data science concepts should contain accurate evaluations of tools, methods used, including data from the validation process and accurate documentation is even more fundamental as strict repeatability becomes very hard to uphold.

3.3 Some tools for tackling the Big Data Challenge

At this stage, due also to the fast-changing landscape in data science, it is hard to systematically categorize its tools and techniques. We review here some of them.

Map-Reduce is a framework used for massive parallel tasks. This works well when the datasets does not involve a lot of internal correlation. This does not seem to be the case for digital evidence in general but a task like file fragment classification is suited to be modelled in a Map-Reduce paradigm. Attribution of file fragments – coming from a filesystem image or from unallocated space – to specific file types is a common task in forensics: machine learning classification algorithms – e.g. logistic regression, support vector machines – can be adapted to M-R if the analyst forgoes the possible correlations among single fragments. A combined approach where a classification algorithm is combined for instance with a decision tree method probably would yeld higher accuracy.

Decision trees and random forests are fruitfully brought to bear in fraud detection software, where the objective is to find in a vast dataset the statistical outliers – in this case anomalous transactions, or, in another application, anomalous browsing behaviour.

In audio forensics unsupervised learning techniques under the general definition of "blind signal separation" give good results in separating two superimposed speakers or a voice from background noise. They rely on mathematical underpinning to find, among possible solutions, the least correlated signals.

In image forensics again classification techniques are useful to automatically review big sets of hundreds or thousands of image files, for instance to separate suspect images from the rest.

Neural Networks are suited for complex patter recognition in network forensics. A supervised approach is used, where successive snapshots of the file system are used to train the network to recognize normal behaviour of an application. After the event the system can be used to automatically build an execution timeline on a forensic image of a filesystem.[KhCY07] Neural Networks have also been used to analyse network traffic but in this case the results still do not present high levels of accuracy.

Natural Language Processing (NLP) techniques, including Bayesian classifiers and unsupervised algorithms for clustering like k-means, has been successfully employed for authorship verification or classification of large bodies of unstructured texts, emails in particular.

4 Conclusion

The challenges of big data evidence already at present highlight the necessity of revising tenets and procedures firmly established in digital forensics. New validation procedures, analysts' training, analysis workflow shall be needed in order to confront the mutated landscape. Furthermore, few forensic tools implement for instance machine learning algorithms or, from the other side, most machine learning tools and library are not suitable and/or validated for forensic work, so there still exists a wide space for development of innovative tools leveraging Machine Learning methods.

References

[BBC+08] Barreno, M. et al.: "Open Problems in the Security of Learning". In: D. Balfanz and J. Staddon, eds., AISec, ACM, 2008, p.19-26

[FBI08] FBI: "RCFL Program Annual Report for Fiscal Year 2008", FBI 2008. http://www.fbi.gov/news/stories/2009/august/rcfls_081809

[FBI10] FBI: "RCFL Program Annual Report for Fiscal Year 2010", FBI 2010.

[ISAC13] ISACA: "What Is Big Data and What Does It Have to Do with IT Audit?", ISACA Journal, 2013, p.23-25

[ISO12] ISO/IEC 27037 International Standard

[KhCY07] Khan, M. and Chatwin, C. and Young, R.: "A framework for post-event timeline reconstruction using neural networks" Digital Investigation 4, 2007

[Pear01] Pearson, G.: "A Road Map for Digital Forensic Research". In: Report from DFRWS 2001, First Digital Forensic Research Workshop, 2001.

[Vari09] Varian, Hal in: The McKinsey Quarterly, Jan 2009