Published in Proceedings of the 2018 International Conference on Computing and Data Engineering, 4-6 May 2018, Shanghai, China, which should be cited to refer to this work.

Weighted Forensics Evidence Using Blockchain

David Billard
HES-SO - Geneva School of Business Administration
Carouge, Switzerland
David.Billard@hesge.ch

ABSTRACT

When digital evidence is presented in front of a court of law, it is seldom associated with a scientific evaluation of its relevance, or significance. When experts are challenged about the validity of the digital evidence, the general answer is "yes, to a reasonable degree of scientific certainty". Which means all and nothing at the same time, since no scientific metric is volunteered. In this paper we aim at providing courts of law with weighted digital evidence. Each digital evidence is assigned with a confidence rating that eventually helps juries and magistrates in their endeavor. This paper presents a novel methodology in order to:

- Provide digital forensics experts with the ability to form a digital evidence chain, the Digital Evidence Inventory (DEI), in a way similar to an evidence "block chain", in order to capture evidence;
- Give experts the ability to rate the level of confidence for each evidence in a Forensics Confidence Rating (FCR) structure;
- Provide experts with a Global Digital Timeline (GDT) to order evidence through time.

As a result, this methodology provides courts of law with sound digital evidences, having a confidence level expressed in metrics and ordered through a timeline. The objective of this work is to add a reliable pinch of scientific certainty when dealing with digital evidence.

CCS Concepts

• Security and privacy → Privacy-preserving protocols • Security and privacy → Usability in security and privacy • Social and professional topics → Computer crime.

Keywords

Digital forensics, digital evidence, e-evidence, blockchain technology, legal evidence admissibility, data provenance.

1. INTRODUCTION

This work aims at concurring to the work of justice by enlightening court rulers and parties about the confidence they can expect from digital evidence. The work takes advantage from new advances in block chain technology and cryptography in order to provide digital forensics investigators with tools to collect and produce e-evidence with associated metrics.

During the process, the forensic practitioner builds three data structures:

(1) The *Digital Evidence Inventory (DEI)*, based on a "block chain" technology, in order to capture evidence. This DEI is immutable and can be used by all parties in a case. Each party has access to the same knowledge about the digital evidences.

(2) The *Forensics Confidence Rating (FCR)* structure. With the FCR, the practitioner grades the e-evidence, based on a categorization of data and data provenance. This rating is subject to modification, depending on the unfolding of the case.

(3) The *Global Digital Timeline (GDT)* to order evidence through time. It is the experience of the author that magistrates and lawyers are particularly sensible to the order of events. It is of utmost importance for the forensics practitioner to provide them with a timeline composed of e-evidence.

The paper is structured as follows: after some related works on the measurement of evidence, we introduce a simple example that will help understanding the inner working of the data structures, that are presented in turn: (1) the *Digital Evidence Inventory (DEI)*, (2) the *Global Digital Timeline (GDT)* and the *Forensics Confidence Rating (FCR)* structure in a more extensive manner. We then conclude this paper with the works in progress.

2. RELATED WORKS

To the best of our knowledge, few literature has exposed a framework that, at the same time, is usable by the expert to characterize the e-evidences and by the courts to base their judgements on facts with a measurable degree of certainty.

One of the most accomplished work in this area can be found in [3]. It follows the lessons learned from the Daubert case [4] concerning the generally accepted guidelines for evaluating scientific evidence that include quantifying the technique's potential rate of error, and the work from Judge Pollack [7] calling for more rigorous requirements. In [3], the author voices the opinion that forensic examiners have a duty to estimate how closely the measured values represented in their data approximate reality.

In a prospective essay on the future of digital forensics, [6] emphasis the fact that the research community should work to develop digital forensic techniques that produce reportable rates for error or certainty when they are run.

If we leave for a moment the digital world to the physical world, most authors and in particular [11] state that evidence admissibility should be determined on the basis of the reliability and accuracy of the process involved.

ICCDE 2018, May 4–6, 2018, Shanghai, China

ACM ISBN 978-1-4503-6393-8/18/05.
https://doi.org/10.1145/3219788.3219792

Most work rely on the validity of the *process*. Although important, most of the processes at the origin of e-evidence are unavailable for analysis. Either because they are unknown, or simply because the source code of the software governing the data creation is closed, or too complex to analyze.

But two main aspects can be detailed:

(1) characterizing how the e-evidence was *collected* and

(2) what *measure of its relevance* and *confidence* one can tag to the e-evidence.

As a matter of fact, battles in courts seldom question the existence of the e-evidence, but rather the reason of its existence. For instance, if e-evidence includes pedo-pornographic images, the debate will focus on why the images were there, with defendants usually incriminating viruses, advertisements on a web page that the suspect did not volunteer, etc.

One can note that e-evidence can also be the absence of data. For instance, when the system log files have been voluntarily deleted from a computer.

The next section presents the data structures needed to store gathered e-evidence, to associate measure and to compose a timeline.

3. DATA STRUCTURES

3.1 Example

In order to explain the data structures devised in this paper, we take a small example of e-evidence, in table 1.

Table 1. Example of e-evidence

Serial #	Name	User	Last connection	First connection
42014287	S3300		04.11.2016 08:52:50	
7299803F	Kingston Data-Traveler 2.0 USB Device	BadGuy	08.11.2016 12:30:11	2016.05.17 12:45:57
182127000	USB Flash Mem-ory USB Device	BadGuy	18.07.2016 12:15:16	2016.07.18 08:39:50

This e-evidence is taken (and modified) from a real case. It is a list of USB devices connected to a computer. This list comes from the USBSTOR Windows registry hive.

3.2 Digital Evidence Inventory (DEI)

The Digital Evidence Inventory is used to capture e-evidence inside an immutable blockchain and forms a traceable e-evidence bag.

The DEI is based on the Scrybe Provenance Framework [9]. In Scrybe, the authors present the components necessary to preserve "provenance data", which means preserving how data was derived. In their work, the authors define a model based on *blockchain technology*, with a *lightweight mining* and *distributed consensus*.

The lightweight mining is achieved with a rapid and small footprint algorithm that can be summarized as follows:

(1) Each miner, from a total of *N* miners, generates a random number.

(2) Each miner broadcasts the hash of their random number.

(3) Once all hashes have been broadcasted, each miner broadcasts its own random number.

(4) Each miner verifies the hashes and calculates

$$Elected\_miner = sum \% N$$

where *sum* is the sum of all the random numbers

(5) The miner with an *id* equal to *Elected_miner* creates the new block and broadcasts it to all other miners.

The proposed model fits our need for the DEI, since we can map almost directly all our components:

- The *transaction* represents a digital evidence that is linked to a case. The provenance of the digital evidence includes at least a case identification, time of acquisition and technical details. The digital evidence itself is represented by a hash (or multiple hashes) of its content, and its location. In some cases, when the evidence is small in size, it can be directly stored into the blockchain in XML format. By the model, the user's cryptographic signature is added to the transaction and therefore indicates the investigator identity.
- The *block* is a collection of transactions.
- The *miners* are the digital forensics investigators working in the same laboratory or office. Contrary to the bitcoin miners, they don't need to present a proof of work, since there is nothing to gain out of mining. The "reward" for mining is to obtain an immutable blockchain.

The version number that was part of the bitcoin transaction is removed in [9], although we think it might be valuable, should the protocol adapts or specializes.

Table 2 shows the differences between the blockchain used for the bitcoin protocol and Scrybe.

Table 2. Model comparison

Component	Bitcoin protocol
Transaction	Proof of present ownership Value oriented
Block	Difficulty and nonce
Mining	Resource intensive
Security	Based on computational difficulty
Component	Scrybe protocol
Transaction	Proof of past ownership Data oriented
Block	Cryptographically signed
Mining	Simple selection algorithm
Security	Based on cryptographic signature

Every e-evidence is recorded into the blockchain on the form of XML tokens. For instance, table 3 presents a simplified view of the blockchain records associated to our example. Each record has additional information, like the device examined, the investigator id, etc. Each record is linked to the previous record and possesses its hash. The blockchain is restricted to a case. Its

peculiarity is that only one miner (the forensics practitioner) adds transactions and all the other miners act only as validators.

Table 3. Simplified blockchain records

<EVIDENCE>
<TransactionID>0001</TransactionID>
<USBSTOR>
<Serial Number>42014287</Serial Number>
<Name>S3300</Name>
<LastCon>04.11.2016 08:52:50</LastCon>
</USBSTOR>
</EVIDENCE>

<EVIDENCE>
<TransactionID>0002</TransactionID>
<USBSTOR>
<Serial Number>182127000</Serial Number>
<Name>USB Flash Memory USB Device</Name>
<User>BadGuy</User>
<LastCon>18.07.2016 12:15:16</LastCon>
<FirstCon>2016.07.18 08:39:50</FirstCon>
</USBSTOR>
</EVIDENCE>

<EVIDENCE>
<TransactionID>0003</TransactionID>
<USBSTOR>
<Serial Number>7299803F</Serial Number>
<Name>Kingston DataTraveler 2.0 USB Device</Name>
<User>BadGuy</User>
<LastCon>08.11.2016 12:30:11</LastCon>
<FirstCon>2016.05.17 12:45:57</FirstCon>
</USBSTOR>
</EVIDENCE>

The validation of the blockchain is not part of the scope of this paper.

3.3 Global Digital Timeline (GDT)

The *Global Digital Timeline* (GDT) is a data structure associated to the DEI. It is a simple key-value database, where the key is a date, or more precisely a timestamp, and the value is a pair consisting of:

- a *reference* to an evidence (a transaction, in the blockchain terminology) in the DEI;
- a *label* tagging the evidence.

A key is not unique, since multiple evidence may share the same timestamp.

With this data structure, it is possible to extract rapidly meaningful information for a given period of time and to present the result even to a non-specialist. Table 4 presents the content of the GDT applied to our example.

Table 4. Example of Global Digital Timeline records

Key	TransactionID	Label
04.11.2016 08:52:50	0001	LastCon
18.07.2016 12:15:16	0002	LastCon
18.07.2016 08:39:50	0002	FirstCon
08.11.2016 12:30:11	0003	LastCon
17.05.2016 12:45:57	0003	FirstCon

This table can be ordered by the key, to find contemporary elements, or by the label if only one kind of element is sought. This structure is not immutable and is flexible enough for easy processing. Each item can be checked against the e-evidence referenced in the DEI.

3.4 Forensics Confidence Rating (FCR)

The *Forensics Confidence Rating* (FCR) is also a key-value database, where the key is a pair consisting of:

- a reference to an evidence (a transaction, in the blockchain terminology) in the DEI;
- a timestamp of the time the rating was issued.

The value is the rating associated to the evidence. In our model, the granularity of the rating is the e-evidence (the transaction).

The rating is calculated thanks to a taxonomy, primarily proposed by the security researcher Bruce Schneier. In [10], he defines a taxonomy of social networking data, and that taxonomy can be extended to fit any digital artifact in or out cyberspace.

Bruce Schneier defines six data types in the framework of social networking and we added three more data types to capture a broader set of e-evidences.

The first six data types defined in [10] are the following:

(1) *Service data* is the data you give to a social networking site in order to use it.

(2) *Disclosed data* is what you post on your own pages, or social media.

(3) *Entrusted data* is what you post on other people's pages.

(4) *Incidental data* is what other people post about you.

(5) *Behavioral data* is data the site collects about your habits by recording what you do and who you do it with.

(6) *Derived data* is data about you that is derived from all the other data.

The three additional data types are the following:

(7) *System data* is produced when a system is recording an action (or lack of action).

(8) *Private data* is basically any data that you don't want to disclose, or only if the person you disclose the data to is at your highest degree of confidence.

(9) *Leaked data* is a mutation of all others types of data (and chiefly private data), over time, to disclosed data. With the added particularity that this data was not supposed to be disclosed by its rightful owner.

Each data type has a confidence rating that is primarily applied to the e-evidence. This confidence rating may evolve, especially when additional elements of the investigation appear. The e-evidence itself may change its type over time (and thus, the associated rating). For instance, a service data can be disclosed and therefore, its initial rating will evolve.

3.4.1 Service data

Service Data may have a high confidence rating since it is usually checked against public records, or via state issued documents. For instance, a bitcoin trading platform will require an official identification document to prove who you are, and to conform to bank-related laws. Of course, identification

documents can be forged, but the suspect is then under the threat of not being accepted on the platform. And the suspect needs to use a platform. Eventually, some information is necessary correct, in order to cash money or to receive goods.

An example of a real case can be found in [2]. In this US district court of Northern District of California, a federal court order allows the Internal Revenue Service (IRS) to order Bitcoin exchange platform Coinbase to give up their customers' identities.

Note that getting access to service data might be hindered by applicable laws, depending of the ruling courts where the service is registered.

3.4.2 Disclosed data

Disclosed data is heavily used by law enforcement or intelligence bodies. Since the data is meant to be read by other people or machines, every dis- closed data of an individual is scrutinized by law enforcement when he/she becomes a suspect. Law enforcement is not the only body interested in disclosed data: lawyers, journalists, insurance companies, activists or common people can access the information.

An example of disclosed data is the bitcoin blockchain, where every transaction is recorded. Provided that you know the public bitcoin addresses used by a person, all his/her transactions can be scrutinized.

Another example, from a real case, is taken from [1] and [5], where a comment left on MySpace is recognized as a central evidence.

M. Clark was found guilty of murdering a two-year-old girl left in his care and was sentenced to life in prison without parole. On appeal, Clark argued that the trial court improperly admitted evidence from his MySpace account in violation of Ind. R. Evid. 404(b). Taking up the "novel question" of the propriety of admitting such evidence, the Supreme Court of Indiana ruled that the trial court did not err in admitting the evidence, particularly where Clark's own testimony made his character a "central issue" of his defense. The verdict and sentence were therefore affirmed.

The confidence rate that can be associated to disclosed data is not as high as with service data.

3.4.3 Entrusted data

The suspect has intentionally left data at someone else social media. In our opinion, this data type is exactly the same, from a forensics point of view, as disclosed data. The only exception being that data is left privately at someone else social media, who in turn makes it public. For instance, a Twitter direct message that is sent to the public tweet list.

As a matter of fact, US courts make strong distinction between private and non-public data [8] (although most of the cases are civil litigations, the concept is the same for criminal justice):

Litigants continue to believe that messages sent and posts made on their Facebook pages are "private" and should not be subject to discovery during litigation. In support of this, litigants claim that their Facebook pages are not publicly available but, instead, are available only to a limited number of designated Facebook "friends." Courts consistently reject this argument, however. Instead, courts generally find that "private" is not necessarily the same as "not public." By sharing the content with others - even if only a limited number of specially selected friends - the litigant has no reasonable expectation of privacy with respect to the shared content. Thus, the very purpose of social media - to share content with others - precludes the finding of an objectively reasonable expectation that content will remain "private."

Provided that law enforcement knows the social media identification code of the suspect, all entrusted data is search-able and may produce an interesting confidence rate.

3.4.4 Incidental data

If one considers *Incidental data*, the order of confidence should be lower than service data, since people can lie or make assumptions, mistakes or alternative truth as it is now called. It can, however, color the suspect profile. For instance, think about a political forum arguing about a candidate to some election.

The rating may not be very high in the first place, but a photo where a suspect appears, or a list of a meeting attendees including the suspect's name, all this constitutes incidental data that may rise the confidence rating.

3.4.5 Behavioral data

For Bruce Schneier, who designed this taxonomy in the context of social media, behavioral data is the information, at large, that a social media can affix to your identity. If you get past the context of social media, we can consider behavioral data as any data that is produced by a legitimate user (human or machine) at any cyber service.

For instance, the list of whatsapp calls you made is be- havioral data. Getting access to behavioral data might be hindered by applicable laws, depending of the ruling courts where the service is registered.

When behavioral data is recorded by an autonomous ma- chine, it can possess a high confidence rating.

3.4.6 Derived data

Derived data is what is produced by artificial intelligence, big data, data mining and such services. Derived data is also what is produced more prosaically as basic oriented graphs showing relations among people or machines. This data is much more in demand in investigations, even for intelligence gathering.

The confidence rating is of course function of the confidence rating associated to the original data.

3.4.7 System data

System data is produced when a system is recording a suspect action (or lack of action). For instance, a suspect enters a room and a sensor is triggered: this is a system data. Another example is when a suspect accesses his/her email box: a system data is produced. Another example is when the suspect phone is Bluetooth enabled and it comes close to another Bluetooth enabled phone: both phones may produce system data with the Bluetooth address and name.

In our example, the e-evidence is taken from a registry hive and is considered of System data type. It can be associated with a high confidence since few people can modify registry keys on a Windows System.

Table 5 shows the confidence rating associated to our example e-evidence. The first e-evidence has no user name associated with the USB key description, so we can lower its confidence rating since we don't know which user inserted the USB key. Thus, creating a new record in the FCR, as shown in table 6.

Table 5. Example of Forensics Confidence Rating records

TransactionID	Time of Rating	Rating
0001	25.12.2017 01:00:00	High
0002	25.12.2017 01:00:00	High
0003	25.12.2017 01:00:00	High

Table 6. Addition of a FCR records

TransactionID	Time of Rating	Rating
0001	25.12.2017 01:00:00	High
0002	25.12.2017 01:00:00	High
0003	25.12.2017 01:00:00	High
0001	25.12.2017 12:30:00	Medium

3.4.8 Private data

Private data is basically any data that you don't want to disclose, or only if the person you disclose the data to is at the highest degree of confidence. For instance, medical records, intimate diary, bank account credentials. During an investigation, all this data might be disclosed to digital forensics investigators, even if they are not related to the case.

The confidence rating of the case related data is usually unknown.

3.4.9 Leaked data

Leaked data is a mutation of all others types of data (and chiefly private data), over time, to disclosed data. With the added particularity that this data was not supposed to be disclosed by its rightful owner.

It is a special case of data because not all courts accept stolen or leaked data as evidence. A good example of these kind of data is the banking records stolen at some financial institutions.

Even derived data can be leaked. Since a tier is involved in the leakage, leaked data can be tampered with before they are released. Therefore, the confidence rating should be low.

4. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a preliminary framework for building a fact-based confidence rating of e-evidence.

First, e-evidence is collected inside an immutable e-evidence blockchain, the *Digital Evidence Inventory* (DEI). Every party in a trial can have access to the DEI that provides also traceability. The blockchain enforces the "chain of evidence": who obtained the evidence; where and when the evidence was obtained; who secured the evidence; who had control or possession of the evidence. In this manner, any modification of the evidence can be traced back.

Then e-evidence is categorized into basic data types and each data type is associated with a measure of the certainty and relevance of the e-evidence. This measure can evolve through time, as well as the categorization. That is the reason why the measures are kept in an external data structure, the *Forensics Confidence Rating* (FCR), linked to the DEI. All the rating modifications are recorded. Each party involved in a trial can have its own FCR, with its own ratings, depending on its own view of the trial.

And finally, a *Global Digital Timeline* (GDT), also linked to the DEI, is created. It is our experience that magistrates are primarily interested in the chain of events that led to a specific crime, or action. This chain of events gives the landscape of the actions taking place before, during, and after a crime is committed.

In overall, this framework allows for a better confidence rating of e-evidence, both for the forensics investigators and the courts. The framework is a valuable documentation tool for the forensics investigators, that can be cross-examined.

Future works include a finer tuning of the blockchain protocol, a semi-automated tool for the building of the GDT and a more precise confidence rating by adding error rate probabilities and relevance.

5. ACKNOWLEDGMENTS

Our sincere thanks to HES-SO for supporting this research partially through the grant ISNet 74354 and HEG Genève for the funding of innovative cybersecurity lab.

The author thanks the anonymous reviewers for their helpful comments.

6. REFERENCES

[1] *Ian J. CLARK v. STATE of Indiana*. (Oct. 2009).

[2] *John Doe vs USA*. (Nov. 2016).

https://assets.documentcloud.org/documents/3228213/Order.pdf

[3] Eoghan Casey. *Error, uncertainty, and loss in digital evidence*. International Journal of Digital Evidence 1, 2 (2002), 1–45.

[4] Chief Justice Rehnquist. *Daubert v. Merrell Dow Pharmaceuticals, Inc*. (1993).

https://supreme.justia.com/cases/federal/us/509/579/case.html

[5] Electronic Discovery Law. *Indiana Supreme Court Rules Trial Court Properly Admitted Evidence of Defendant's MySpace Page in Murder Trial*. (Oct. 2009).

https://www.ediscoverylaw.com/2009/10/

[6] Simson L. Garfinkel. *Digital forensics research: The next 10 years*. Digital investigation 7 (2010), S64–S73.

[7] Judge Pollack. *UNITED STATES of America, v. Carlos Ivan LLERA PLAZA, Wilfredo Martinez Acosta, and Victor Rodriguez*. (March 2002).

https://law.justia.com/cases/federal/district-courts/FSupp2/188/549/2576958/

[8] Margaret (Molly) DiBianca. *Discovery and Preservation of Social Media Evidence*. (Jan. 2014).

https://www.americanbar.org/publications/blt/2014/01/02_dibianca.html

[9] Ujan Mukhopadhyay, Carl Worley, Anthony Skjellum, Oluwakemi Hambolu, Xingsi Zhong, Jon Oakley, Lu Yu, and Richard Brooks. 2017. *The Scrybe Provenance Framework: Scalable Secure Data Provenance Using Blockchain Technology*.

[10] Bruce Schneier. *A Taxonomy of Social Networking Data*. IEEE Security & Privacy 8, 4 (2010), 88.

[11] Strong, J. W. *McCormick on Evidence*. 4th Ed. section 294 (1992)