MEASURING INCONSISTENCY METHODS FOR EVIDENTIARY VALUE

Fred Cohen
California Sciences Institute
California, USA

Many inconsistency analysis methods may be used to detect altered records or statements. But for admission as evidence, the reliability of the method has to be determined and measured. For example, in China, for evidence to be admitted, it has to have 95% certainty of being correct,[1] and that certainty must be shown to the court, while in the US, evidence is admitted if it is more probative than prejudicial (a >50% standard).[2] In either case, it is necessary to provide a measurement of some sort in order to pass muster under challenges from the other side. And in most cases, no such measurement has been undertaken.

The question of how to undertake a scientific measurement to make such a determination, or at least to claim such a metric, is not well defined for digital forensics, but perhaps we can bring some light to the subject this issue.

CAUSALITY, REFUTATION, AND CONFIRMATION BIAS BASICS

I have said and will likely say this often and again. Effect does not imply cause. Rather cause (C) acting through mechanisms (m) produces effects (E), expressed as $C \to^{m} E$. To have a scientific hypothesis, it is not enough to state that C produces E, it is also necessary to identify the mechanism by which C produces E. Testing can then be repeatedly done to confirm or refute the hypothesis of $C \to^{m} E$ by trying to refute the hypothesis. If refutation fails, it is a confirmation, while if refutation succeeds, the hypothesis as stated cannot be correct. While many confirmations may be found, any number of confirmations of a universal statement do not prove it to be correct, while a single refutation demonstrates its falsehood.[3] Typically, science progresses when a refutation is identified, the errors in the $C \to^{m} E$ hypothesis are identified, and an updated $C' \to^{m'} E'$ version of the hypothesis is created to mitigate the refutation cases, or the hypothesis is abandoned.

Studies over many years show that prior beliefs or information tend to effect outcomes and, specifically, that scientific results are biased toward prior beliefs.

[1] Zhang, Xiao, Chinese graduate student in law. Personal correspondence.

[2] Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 US 579, 125 L. Ed. 2d 469, 113 S. Ct. 2786 (1993).

[3] Popper, K. *The Logic of Scientific Discovery* (1959). London, England: Hutchins and Company. ISBN 10: 0415278449.

This is often called confirmation bias.[4,5] If confirmation testing is done instead of refutation testing, there is a tendency to have false confirmations, and the stronger the prior belief, the more pronounced the confirmation bias effect. This is not malicious, it is human nature, and the nature of science when performed by people.

While we might expect that in a system in which prosecutors control all examination of evidence there would be an extremely high rate of successful prosecution, China has a 30% acquittal rate[6], and only 50% of reported cases involving digital evidence in 2012 went to trial[7], in some part because the 95% certainty required for forensic evidence cannot always be met. A prima facie case could be made for requiring certainty metrics on evidence compensating for confirmation bias, but insufficient evidence is currently available.

INCONSISTENCY ANALYSIS EXAMPLES AND LIMITATIONS

Detecting inconsistencies between records is fundamental to challenging evidence, and as such, it is fundamental to establishing credence. Because sound science is based on refutation rather than confirmation, scientific hypotheses are tested by trying to show they are wrong rather than trying to show they are right. This avoids confirmation bias and establishes that causality is consistent with repeatable testing. Inconsistency analysis consists of various methods for testing records against other records and statements (i.e., traces and events)[8] for type C (internal) and type D (external) consistency. If and to the extent records are inconsistent, they cannot all be true. For example, if a person states that they were at a place at a time, and records show that they were elsewhere at that time, then the statements, the records, or both must be wrong. While the seemingly logical assertion detailed in the footnote[9] has not reached more than random level

[4] Koehler, J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. Northwestern University, School of Law, September 1, 1993, *Organizational Behavior & Human Decision Processes*, *56*: 28.

[5] Darley, John, Gross, M., and Paget, H. (2000). A hypothesis-confirming bias in labeling effects. In Stangor, Charles, *Stereotypes and Prejudice: Essential Readings, Psychology Press*, p. 212, ISBN 978-0-86377-589-5, OCLC 42823720.

[6] Zhang, Xiao, Chinese graduate student in law. Personal correspondence.

[7] First keynote speaker at First International Summit of Electronic Evidence and 2013 China Forum of Cyber Crime and Social Security, 2013-05-23, Shanghai, China.

[8] Cohen, F. (2009). Two models of digital forensic analysis. IEEE/SADFE-2009, *Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, in conjunction with the IEEE Security and Privacy Symposium Oakland Conference, Oakland, CA, USA, May 21, 2009.

[9] Definition: Consistency between two or more things means that each is the way you would expect it to be if the other ones are the way you observe them to be. Example: If you see a black box and someone else viewing the same object under the same conditions states that it is a white sphere, your observation is inconsistent with their

of consensus (68% agreement, which is inadequate for consensus at the sample size as of the last study on this issue), we will assert it nonetheless.[10]

In most digital systems, many different and redundant traces are produced, and thus many different tests for inconsistency may be found. Examples include: (1) travel information with computer and facility access logs, which should be consistent in terms of time and location; (2) computer login and online credit card usage, which should indicate that the user was logged in when their credit card was charged for an online transaction; (3) DNS, Web access, and email records, which should indicate DNS logs just prior to uncached Web access and MX record lookups just after initiating emails; and (4) identity management system records and access to department databases, which should indicate that those accessing records from a given department have roles matching the records accessed.

But just because two or more records appear to be inconsistent doesn't actually mean that either or any are wrong. In fact, records may indicate seemingly inconsistent situations and all be right, in the sense that they are all effects of consistent causes acting through identifiable, but different, mechanisms. To get a sense of this, suppose we have log files, file system state, and user statements in which the stated time sequences for the same acts took place in different orders (e.g., (A, B, C), (A, C, B), and (C, B, A)). This would seem, on its face to be inconsistent. But as we examine causality of the actual mechanisms at play, we might identify that the different mechanisms work in different ways and thus produce records in different sequences or with different precision and accuracy.

- A person may indicate arrival at 08:00, login at 08:05, and file access at 08:10;
- while the badge reader indicates entry at 08:05, the login record on the computer indicates login at 08:04:31, and the file server indicates file access at 08:03:25.2316;
- and the log server may record door entry at 08:02:12, file system access at 08:03:18, and login at 08:04:15.

statement. Similarly, if a sworn statement states that a particular file was created at 10AM on a particular day in a particular place, and the metadata for the file indicates that the same file was created at a different time on a different day, the sworn statement is inconsistent with the metadata. Assuming this definition as the basis for your answer, respond to the following statement: As a fundamental of digital forensics, what is inconsistent is not true (or in other words, the inconsistent things cannot all be true).

[10] Cohen, F. (2012). Update on the state of the science of digital evidence examination. *Conference on Digital Forensics, Security, and the Law*, 2012.

The only way to really understand these records properly is by experimentation, which we call reconstruction. In this process, we reconstruct a similar environment under various conditions and seek ways in which the event sequences may be reconciled. In the process, we likely will learn about the precision and accuracy of the different time sources and recording methods and about links between cause, mechanism and effect. If we perform enough experiments to determine that the records cannot be reconciled, then we have a definitive refutation in the sense that the set of claims from the three sources cannot all be true; another hypothesis about how they came to be must be sought. But if we find a reconstruction that shows similar records and sequences, we may refute the definitive claim of inconsistency, yet this does not prove the records to be right. Indeed we can almost never prove the records right because we cannot normally exhaust all possibilities.

As an alternative to reconstruction, which is often expensive, we may do analysis to refute hypotheses. For example, we may take a look at historical records over the same time period and identify that other cases of different orderings between records of the log server and badge reader, login records, and file server access are found. This would explain the apparent inconsistency. But if no such instances are found, that only confirms the hypothesis that the records are normally consistent in this respect and does not explain the causal chain associated with the events in question. It does not prove altered records or demonstrate that the person is lying, and it does not provide a causal chain explaining apparent inconsistency.

SOME PROBLEMS WITH COMPUTER-RELATED RECORDS

The truth of computer-related records is that they often appear to indicate things that did not in fact happen. That is, effects appear to indicate causes, but in fact the apparently indicated causes may produce other effects and other causes also produce undistinguished effects. This is because the mechanisms presumed and apparently indicated are not really the mechanisms present.

> **Example**: The apparent and often presumed indication of a record stating that a person entered a door at 08:05 is that the person actually entered the door at that time. But the actual mechanisms that produce those records may be based on clocks not perfectly set and of limited resolution, and are associated with unlocking mechanisms and not movements of people. The methods by which they are produced and recorded may have delays, end up in different orders, and reflect different physical mechanisms than the outputs seem to indicate.
>
> **Explanation**: The recording mechanism of the badge reader system keeps records at 5 minute precision rounded up when it sends logs of those records to the logging server as the badges are read. Thus the badge reader record of 08:05 is consistent with the logging server record of

> 08:02:12. The human recollection of time has limited precision and accuracy, may be based on a wrist watch or other clock besides the badge reader, and may be recalling the arrival at work rather than use of the badge reader.

We now have a possible causal chain that is consistent in terms of facility entrance. Let's look at logins next:

> **Example**: The record of computer login at 08:04:31 and the log server record of computer login at 08:04:15 are consistent if we assume that clock accuracy is only good to the nearest minute. But assumption is the problem here. We would need to examine further to determine if this is in fact the case. Suppose it is not, and that both systems use a common time base and are measured regularly so that we can demonstrate that they are synchronized to within one second at almost all times and that other sorts of records contemporaneous with those at issue are within a second of each other.
>
> **Explanation**: It turns out there are other ways these times could differ by 16 seconds. For example, it might be that the login record in the logging server reflects a timestamp collected at the start of the login process, while the login time recorded on the computer system reflects the time at which the password was determined correct. These could easily be 16 seconds apart. Of course the human time remains too inaccurate to be distinguishable and we are now consistent again.

But what of the file server access records?

> **Example**: The file system records indicate access at 08:03:25.2316 while the logging server indicates access at 08:03:18. Obviously, there is a difference in precision of the records displayed. But in addition, the logging server seems to say that the access occurred before the file system indicates it. The synchronization issue returns, as does the issue of what actual process produces which records.
>
> **Explanation**: Suppose the file system sends an audit record to the logging server when a file is opened but records last access only when the file is closed. Then the records would be consistent with a file open, access and close over a 7-second period, even if the clocks were closely synchronized.

But what of the sequence? Even if each redundant record may be independently reconcilable, the overall sequence seems infeasible, and inconsistent across explanations. How can a user access a file before being logged in or present, and how come the sequence identified by the user differs from that of the computer records?

Of course it turns out that there are plenty of feasible explanations for the differences in sequence as well. But none will be provided here for this example. The problem is that all of this is speculation, as is any claim that the records are inconsistent or consistent, unless and until there is a scientific basis for making claims of causality based on the traces produced as effects.

MEASUREMENT THEORY AND REACHING A DEFINED LEVEL OF CERTAINTY

The resolution of this issue must ultimately come in the form of a theoretical basis for making and using measurements and actual measurements tested against that theory. Today, we don't, as a field of scientific endeavor, have a widely agreed-upon theory of measurement for digital traces.

> ***It's not that we disagree all the time. As a community, those who work in digital forensics largely agree at some poorly-defined level of consensus. Perhaps the biggest problem we face is that we often agree on things that are demonstrably wrong.***

It is common practice today to make timelines based on digital records such as those detailed above, and to try to produce claims of causal sequences based on the effects reflected in those traces. But in almost all cases I have seen, there is no initial determination that the time-related records accurately and precisely reflect things that can be ordered relative to each other at the level of time differences at issue based on those records or the mechanisms that produce them. In simpler terms, the claims are made without adequate basis. In most cases, no basis is provided at all, and the time-related records are assumed accurate and precise as shown.

To get a sense of how far this is from a good solution, note that even if a basis was provided, that basis would have to be shown reliable to a level of certainty appropriate to the evidence admissibility requirements at hand. To date, I have rarely seen an instance where it was even demonstrated that a record was differentiable from a random number in any probabilistic sense. The business records exception is commonly used to admit evidence, to wit: “this record is regularly collected in the course of business and relied upon for its day-to-day operations.” This establishes an unknown reliability and non-scientific basis for admission.

I am not a statistics expert, but, at a minimum, to reach 95% likelihood that a cause produces an effect, an experiment would have to be repeated 20 times with no more than one case in which the cause failed to produce the effect. While this is a starting point, sampling theory tells us about standard distributions and random samples, but is typically not probative with respect to the highly repeatable and brittle processes associated with the finite state automata that produce traces used as digital evidence.[11] In order to link a cause through a mechanism to an effect, it might, for example, be useful to be able to turn the mechanism off and back on, producing no effect with the mechanism off and the identified effect with the mechanism on. This is realistically attainable as a reconstruction for the instances above involving digital records. For example, here is a starting point:

- **The badge reader indicates entry at 08**:**05**: Check the time records produced by using the badge reader, first under existing conditions, then under other conditions, including enabling and disabling the logging mechanisms, setting times in its internal clock to different values, and entry during different times relative to the start and end of an hour (e.g., enter at one minute intervals, offset by 10 seconds for each experimental run, producing 6 runs for 0, 10, 20, 30, 40, and 50 seconds after each minute). This might produce enough records to be able to demonstrate the precision and accuracy of the badge reader internal recording mechanism to the desired level of certainty and test the hypothesis of round up, time offset errors, etc.
- **The log server records door entry at 08**:**02**:**12**: Repeat the above experiment with the log server enabled and not enabled, with the standard initial time of the log server and variations on time settings, and under various load conditions of the server and network. The idea is to see whether and to what extent differentials in time may reflect different mechanisms that might produce them, and then based on these results, look for indicators of those mechanisms in the relevant time frames to the issues at hand. While, ideally, we would produce an actual entry at 0800 with badge reader and audit records producing identical values to those in the actual case, demonstrations of the level of variation and conditions should be adequate to demonstrate reliability of those records relative to each other, or in the alternative, failure to observe any such variations under any identifiable test conditions should indicate that the various hypotheses are refuted by the tests.
- **The login record on the computer indicates login at 08**:**04**:**31**: Just as for the cases above, and in particular for identified cases such as the difference between timestamp collection and recording times and phases of the login processes, testing can be performed to determine how closely correlated physical activities and recorded records are for each of the relevant records as well as differentials and conditions which cause them between computer system logs and logging server logs. Hypotheses about what logs are produced by what mechanisms at what

[11] Cohen, F. (2012). *Digital Forensic Evidence Examination*, 4th ed. ASP Press, 2012.

times can be tested by enabling and disabling various mechanisms, to the extent that is feasible.

- **The file server indicates file access at 08:03:25.2316**: This has similar testing requirements to the records produced by logging in. In the case of file servers, things can get more complicated, for a wide range of reasons. Some such reasons include different clock and recording resolutions for different file systems (e.g., FAT file systems retain time stamps to a 2 second resolution (which means a 4 second maximum range for a given time) while NTFS counts at 100 nanosecond intervals and EXT3 is 1 second, while EXT4 is 1 nanosecond[12]), multiple time stamps (e.g., modify/access/create may be produced by different mechanisms in different ways), and I am unaware of any file system that records times to 4 digits of precision, an apparent inconsistency with the data provided above.
- **The log server recorded door entry at 08:02:12, file system access at 08:03:18, and login at 08:04:15**: Without a lot more detail about the mechanisms in use, a wide range of possibilities exist. For example, we would want to differentiate the mechanisms by which times might be recorded and information sent to the logging server. Some records might be batched and sent periodically or when a buffer is filled, while others might be sent in real-time. Some might be sent via the UDP protocol, which drops datagrams but doesn't typically delay them significantly, while other mechanisms might use TCP which has retries and other reliability features. Does the logging server records when it gets records or rely on timestamps from those records? Does it always add new records to the end of a log file or does it have multiple log files? Is the order of recording the same as the order of arrival? Is there a prioritization mechanism in the network or logging server? What happens as it runs out of disk space? The list goes on.

Reconstruction is non-trivial and, to be efficient and effective, one must select what to test in what way and with how many samples in order to produce reliability information adequate for admission. It cannot generally be based on historical testing in “similar” systems. In fact, reproducing behavior often comes down to the exact set of software and revisions of an operating environment, and I have had cases where this made the difference in outcomes in the legal matter.

Many experts merely claim that various things are true based on their experience, but the expert who wants a scientific basis should test such claims in situ to determine whether they are refuted. In case after case, common wisdom and asserted experience has been refuted by reconstruction.

[12] Need citation(s) for these times.

Why not reverse engineer instead of reconstructing, etc.?

The simple answer in the US is that it's illegal[13]. For "computer security" reasons, reverse engineering is legal as of a Federal Trade Commission ruling of a few years ago, but for forensics purposes, it is still not permitted. So the legitimate expert cannot look at a disassembled version of the binary of the operating system to determine how a record is produced.

This is similar to many other things in forensics. While there are many ways to address a technical question, legal systems limit what can be done and how they can be done. As a scientist, you might like to do anything you like, but as an expert witness in a legal matter, you cannot just try things. Actions are limited by the client's willingness, the court's desire, available time and money, what is already known science, and the legal process you work under.

Even if we could reverse engineer all of the software and hardware involved in typical systems of today, environmental conditions also have an effect on records produced. High load conditions, lack of storage space, and other extremes in resource availability have pronounced effects. But far more subtle causes, such as single bit errors in memory, interactions between multiple processes and programs, unanticipated input sequences, and a long list of other similar causes[14] may produce changes in the apparent operations of interacting mechanisms. The state of the art in software and hardware analysis is inadequate to effectively predict the envelope of effects from the reverse engineered mechanisms or to rule out other potential mechanisms.

EDITORIAL SUMMARY AND DISCUSSION

Building the science of digital forensics is fundamental to advancing justice in the information age. We are at a unique point in time where we can make an enormous difference in this advancement. The field is emerging, science in legal matters is under scrutiny, and tie information age is bringing about a dramatic emergence of digital evidence in the courtroom. Science is not local to a jurisdiction, and effective law enforcement in the information age requires global cooperation. Around the globe, scientists are seeking to make progress, and they are succeeding, albeit slowly and with little support.

Finding ways to measure claims about evidence in legal matters is not a simple or even well-understood matter. Reconstruction can often be done under

[13] Cohen, F. (2010). The DMCA Still Restricts Forensics. Retrieved from http://http://all.net/Analyst/2010-08.pdf on August 1, 2010.

[14] Cohen, F., Phillips, C., Swiler L. P., Gaylor, T., Leary, P., Rupley, F., and Isler, R. (1998). A cause and effect model of attacks on information systems. Some analysis based on that model, and the application of that model for cyberwarfare in CID. *IFIP-TC11 Computers and Security*, *17*(3): 211-221.

conditions anticipated or asserted, and the results of repeated testing under different reconstruction conditions can be revealing in terms of numerical assessments of reliability and refutation or confirmation of various claims. Of particular use is testing with and without mechanisms active, as this differentiates the causal chains associated with those mechanisms and allows claims of causality to be shown, with repetition producing results of the form of "[X] of [Y] of the cases tested demonstrated [causal mechanism] which is consistent with [claim]. This is a [100*X/Y]% reliable result."

I invite all of those reading and writing for this publication and others to engage in the discussion of the advancement of the science, and to do so in the form of letters to the editor.

I look forward to the opinions of others–sent to the editor, to me, or as part of published work.