Two models of digital forensic examination

Fred Cohen, Ph.D.
CEO - Fred Cohen & Associates
President - California Sciences Institute
fc at all dot net

Abstract

This paper examines an existing cost model of digital forensic evidence examination, identifies minor optimization improvements to that model, describes a new model, and uses the new model to show some fundamental theoretical limits of examination.

1. Introduction and background

Digital forensic evidence (DFE) analysis and interpretation is largely rudimentary and ad-hoc, but there is a substantial movement afoot to introduce increased rigor into these process elements. [1][2][3]. Classic forensic science [5] has theories increasingly being applied to digital forensics. One of the defining principles is that of trace evidence. In the classic formulation, the theoretical basis of trace evidence is that when two objects come into contact with each other, each leaves something of itself with the other. A trace of contact may be found, if sought, and depending on the nature of the trace, it may allow classification of the trace as to type, and/or individualization to a particular object. For example, when a person wears an item of clothing, hair, skin, and sweat from the person attach to the clothing, and fibers from the clothing attach to the person, even if at a particle level. If a trace of the clothing, such as a fiber, is found on the person, it may be used to type the clothing as to color, fabric, manufacturer, model number, and so forth. The person might be identified as to type, such as blood type or hair curliness, and with DNA, may be individualized, sometimes to the specific person.

In the computer arena, the theory of trace evidence is applied in the sense that when actions are taken by actors within a computing environment, many records may be produced that trace the activities. For example, there may be audit trails associated with logins and program execution, time stamps on files and within headers, and so forth. These traces may allow classification of actors and actions by type, and in some cases, individualization of various sorts. For example, the presence of carriage returns at the ends of lines might indicate a Microsoft operating environment or particular sort of software in use, and the presence of a MAC address in an Ethernet packet may be used to individualize the activity to a particular network interface card.

Just as physical evidence may be planted, removed, forged, or otherwise altered, so may DFE. As a result, the reliability of evidence in both realms is subject to challenges, and the study of forensic science is, in no small part, about the development of methods to allow the reliability of evidence to be established, demonstrated, and challenged. The most common scientific methodologies are the use of redundancy to show that traces are consistent, the use of refutation to show that traces are inconsistent, accepted publications of methods in peer reviewed articles in the field, and the use of repeatable experimental evidence from prior or controlled instances to demonstrate the soundness of the technique and approach. The legal methods applied to this evidence are based on the existence of specific statutes and precedents to show that the same sorts of evidence have been accepted or rejected in prior matters.

The scientific study of DFE is in its infancy. As an example, in [1] a study of High Technology Crime Investigation Association (HTCIA) members identified steps taken to collect DFE. A review of the 103 steps identified was done by testifying technical and legal experts to identify which steps were (1) absolutely prohibited; (2) undesired; (3) make no contribution and causes no harm; (4) desired; and (5) absolutely essential. The results showed both significant deviation overall and, for some steps, complete disagreement among experts both between and within groups, even to the extremes of some indicating that the same step was absolutely prohibited that others indicated was absolutely essential. The model in this case could be readily characterized as a partially ordered set with 103 different nodes, most linearly related in sequence from the start of the collection to the end with a relatively small (~3) branches at any location and a relatively short maximum branch length (~3) before rejoining the linear sequence.

In [2], an approach to cost-effective digital forensics based on an initial pre-processing step followed by the use of a Bayesian network model was proposed and demonstrated for a sample case. This is further detailed below.

In [3], a fault model is put forth as a way to challenge digital forensic evidence, thus introducing refutation into the analytical process: "This model assumes that digital forensic evidence is identified, collected, transported, stored, analyzed, interpreted, reconstructed, presented, and destroyed through a set of processes. Challenges to this evidence come through challenges to the elements of this process. Faults consist of intentional or accidental making or missing of content, contextual information, the meaning of content, process elements, relationships, ordering, timing, location, corroborating content, consistencies, and inconsistencies. Not all faults produce failures, but some do. While it may be possible to challenge faults, this generally does not work and is unethical if there is no corresponding failure in the process. Certain things turn faults into failures, and it is these failures that legitimately should be and can be challenged in legal matters. Failures consist of false positives and false negatives. False negatives are items that should have been found and dealt with in the process but were not, while false positives are things that should have been discarded or discredited in the process but were not."

Many other models have been in widespread use for a long time, including models used by detectives and lawyers, ranging from simple linguistic depictions that detail out the different legal elements required in order to demonstrate that the requirements in the law have been met with respect to a crime; to formalizations of laws to be enforced by access controls that automatically take into account jurisdictional issues. [4] The application of these models to digital forensic investigation is in its infancy, but they are increasingly being considered and applied to cases as an approach to reducing error rates among analysts and providing increased assurance and, in some cases, documentation of diligence.

2. An existing model

The approach taken in [2] may be characterized as a legal requirement (L:{l1, ..., ln}) associated with a violation (V) consisting of the union of a set of circumstances such that each circumstance must be shown true to within the standard of proof in order to warrant the charge of a violation based on the defined legal criteria.

1. For each element of the legal requirement [$\forall l \in L$] there is a set evidence chains E: {E1, ..., Eo}, each of which consists of a set of events (e) evidenced by any of a set of

traces T:{t1, ..., tn} of those events within the digital system [$\forall Ex \in E, Ex: \exists \{ex1, ..., exp\}, \forall eab \in Ex, \exists tc \in T: tc \rightarrow eab$].

2. Each item of evidence has an assumed weight Wx=(wx1, ..., wxp) normalized to a total weight of 1, so that $\sum(wx1, ..., wxp)=1$, and a cost of detection cxa so that to total cost of detection for any given chain of evidence is fixed $Cx= \sum(cx1, ..., cxp)$.

3. An investigation starts in phase 1, and as the investigation proceeds, each item of evidence detected contributes to the weight and each effort to detect evidence contributes to the cost. If W exceeds an organizationally defined threshold of adequacy (g), the investigation goes to phase 2. If W get low enough that the total available weight of evidence left to detect cannot reach g, the investigation is abandoned.

4. In phase 2, a Bayesian network is used to analyze the evidence against a hypothesis of how the crime was committed. This network uses a-priori probabilities of traces indicating guilt and yields a probability of guilt (G).

5. When $\forall l \in L, \exists Ex: Px > gx$, G is adequately established to propose charges. In [2] G is calculated as the product of the a-priority probabilities. For example, the presence of a known Trojan can be established with a probability of approximately 0.98, if the claim of anti-virus vendors can be believed.

Legal precedent provides well-established subsets of the overall structure, so that the full complexity of the space is not normally exercised. Once a successful prosecution is made, the evidence required for the particular path through the structure is established and the same elements may be repeated with greater certainty of success in court. This increases the weight of the elements of that particular path. The prior development of methods to establish that path through the structure may also be reapplied to reduce the cost of using that same methods to make future cases. The strategy applied in [2] was to use an existing path based on precedence and identify the lowest cost element of the evidence sequence for each step in detection and analysis. In this way, if a required element is not found, a lower cost is expended prior to determining infeasibility, and more most expensive detection is delayed until required. There is also an implicit assumption in this model that elements are independent, costs are independent, and benefits do not accrue across multiple paths. In effect, multiple paths are not typically taken in this approach because the overall value of detecting any particular criminal committing any particular crime is not normally high enough to justify complex examination. Many potential crimes with much evidence is available to consume resources, and resource minimization with conviction maximization is the goal in [2].

The model from [2] also ignores strategies of opponents and the implications of these strategies over time. For example, by looking at precedents and/or understanding the model being applied, an opponent could determine a strategy by which they could (1) increase the cost of detection by just enough to prevent detection with minimal effort on their part, (2) analyze the structure to identify minimum cuts in the evidentiary paths and create tools to severe the paths while still keeping the costs of detection high, or (3) use high valued cases to challenge precedents and, based on overturning a single case, revisit many prior cases.

The model from [2] is also oriented toward the charging party and is thus inconsiderate of the party being charged. While from a law enforcement standpoint in criminal cases this may seem reasonable and prudent, from a standpoint of justice, and from a civil perspective, advantaging the charging party is problematic. But perhaps

more importantly to those who take the charging party only perspective, ignoring challenges provides too little information to the legal team to survive those challenges, cost optimization in detection implies elimination of redundancy which also makes the case brittle in that a successful challenge of even one element of one evidence chain potentially destroys the entire case, and ignoring the interactions of investigative sequences leaves the potential for further optimization untapped.

A case example was presented in [8]. In this case, a party was considered for charges of originating the release of copyrighted material (a video) using "BitTorrent" for distribution. The relevant legal hypotheses was: "H: The seized computer was used as the initial seeder to share the pirated file on a BitTorrent network." This hypothesis is assumed adequate to consider charging a criminal violation of the relevant copyright statute, because it fulfills the elements of the crime as indicated by the sub-hypotheses. {H1, ..., H5} each of which may be supported by a corresponding set of events {E1, ...E5}:
- H1 (E1): The pirated file was copied from the seized optical disk to the seized computer.
- H2 (E2): A torrent file was created from the copied file.
- H3 (E3): The torrent file was sent to newsgroups for publishing.
- H4 (E4): The torrent file was activated, which caused the seized computer to connect to the tracker computer.
- H5 (E5): The connection between the seized computer and the tracker was maintained.

These event sets are comprised of subsets of the events and/or traces identified in Figure 1, each with an organizationally set cost of trace detection (C), an experientially based assigned evidential weight (W), and is either detected (T) or not (F) (D). The link between the sub-hypotheses $\{H_1 ... H_5\}$, events, and weights are asserted [2] as follows:
- H(E): $W(E)=\pi(W(E_1),..W(E_5))$ where π is taken in this case to be the product of sums of the constituent weights divided by the maximum possible weights.
- H_1:$E_1=\{e_1,e_2,e_3\}$, $W_1=\sum(d_1*w_1, d_2*w_2, d_3*w_3)/\sum(w_1, w_2, w_3)$
- H_2:$E_2=\{e_4*e_5*e_6*e_7*e_8\}$, $W_2=\sum(d_4*w_4, d_5*w_5, d_6*w_6, d_7*w_7, d_8*w_8)/\sum(w_4, w_5, w_6, w_7, w_8)$
- H_3:$E_3=\{e_4*e_8*e_9*e_{10}*e_{11}*e_{12}*e_{13}\}$, $W_3=\sum(e_4*w_4, e_8*w_8, e_9*w_9, e_{10}*w_{10}, e_{11}*w_{11}, e_{12}*w_{12}, e_{13}*w_{13})/\sum(w_4, w_8, w_9, w_{10}, w_{11}, w_{12}, w_{13})$
- H_4:$E_4=\{e_6, e_{13}, e_{14}, e_{15}, e_{16}, e_{17}\}$, $W_4=\sum(d_6*w_6, d_{13}*w_{13}, d_{14}*w_{14}, d_{15}*w_{15}, d_{16}*w_{16}, d_{17}*w_{17})/\sum(w_6, w_{13}, w_{14}, w_{15}, w_{16}, w_{17})$
- H_5:$E_5=\{e_{13}, e_{18}\}$, $W_5=\sum(d_{13}*w_{13}, d_{18}*w_{18})/\sum(w_{13}, w_{18})$

For example, if traces of all events are detected with the exception of events e_{13}, e_{14}, and e_{17}, the resulting calculation yields $W_1=1$, $W_2=1$, $W_3=5.5/7.5$, $W_4=3.5/6.5$, $W_5=0.5/2.5$, $W(E)\approx0.08$. Clearly the influence of e_{13} which is included in traces 3, 4, and 5, is critical, and in this case, a trace must be found of "Internet connection is available at the relevant time" or the overall case is clearly at risk. If such a trace is found, this analysis produces, instead, 1*1*1*5.5/6.5*1, so $W(E)\approx0.85$. In [2], cost is proposed as a driver for ordering the forensic process with the notion that cost will be less if the lowest cost event traces are sought first.

e#	Event and/or nature of trace	C	W	D
1	Modification time of the destination file is after its own modification time	1	1	
2	Creation time of the destination file is after its own modification time	1	1	
3	Hash value of the destination file matches that of the source file	1	1	
4	BitTorrent client software is installed on the seized computer	2	2	
5	File link for the shared file is created	1	0.5	

e#	Event and/or nature of trace	C	W	D
6	Shared file exists on the hard disk	1	2	
7	Torrent file creation record is found	1.5	2	
8	Torrent file exists on the hard disk	1	1	
9	Peer connection information is found	2	0.5	
10	Tracker server login record is found	1.5	0.5	
11	MAC time and link file corroborate Torrent file activation time	2	1	
12	Internet history record about publishing website is found	1.5	0.5	
13	Internet connection is available at the relevant time	1.5	2	
14	Cookie of the publishing website is found	1.5	0.5	
15	URL of the publishing website is stored in the web browser	1	0.5	
16	Web browser software is found	1	1	
17	Internet cache record about the publishing of the torrent file is found	1.5	0.5	
18	Internet history record about the tracker server connection is found	1.5	0.5	

Table 1 - Events, costs, and weights from [2]

3. Analysis of this model

Optimization methods have been long studied in the operation research arena, and many papers have been written on optimizing graph traversal with weightings on nodes and links. Variations on the traveling salesman problem have been shown NP-complete and optimal solutions for substantial graph sizes are infeasible. There may be nearly linear graphs, such as those commonly appearing in the digital forensic analysis process arena, for which optimization is attainable in practical time. For example, the methodology for collecting evidence identified in [1] has little redundancy and only one major path with a few alternative paths, each with limited diversity. As such it is easily analyzed for cuts and most steps are linearly additive in analysis of weight. The metrics provided by the study in [1] also provide the means to associate asserted weights based on the sample set. However, the results of [1] model the forensic process of evidence collection and ignore the legal situation, while the model in [2] relates to the legal situation and largely ignores the forensic process. It is potentially problematic to seek optimization without a full picture of the situation. For sequential activities, such as investigations of the sort assumed in [2], a step-by-step approach may be taken so that by picking the proper sequence of steps and stopping the process when it is determined that the total weight of the evidence cannot exceed the required threshold, cost may be saved in investigations that will ultimately not pay off, and redirected toward more promising investigations. Graph and other mathematical optimization techniques may be applied to problems such as the ones identified in [2] for improvements over the "minimum cost first" approach identified therein. In particular, we propose some alternatives as starting points to improvement in cost efficiency under the model of [2].

A common practice in business is the use of return (R) on investment (I) (i.e., ROI) to determine which activities to perform. ROI may be calculated as the return divided by the investment (ROI=R/I). For ROI<1, R<I and investment should never be made. For ROI>1 risk to be considered. Higher R is generally associated with higher risks. For constant R, R/I may vary significantly between strategies. In [2] R is in the form of

evidentiary weight (w) of a particular method to find a trace of an event, and I is the cost of seeking such a trace. Thus ROI for an activity (a) is (w_a/c_a). The model of [2] defines these, and a simple calculation is made for each activity. Results are sorted from highest ROI to lowest, and process ordered based on this calculation. ROI is calculated with commensurable monetary fungible units for R and I and c and w are not directly fungible, monetized, or commensurable, the use of (ROI<1) as a cut-off does not apply. Activities that appear in more than one evidentiary chain produce an effective weight of the activity as the product of w times the number of event sets in which an event exists while c is only counted once because the activity only has to be done once.

e	w	c	n	ROI	Ch	CR	R	R2	RC	RCR
1	1	1	1	1	1.1	1	5	2	6	3
2	1	1	1	1	1.1	1	5	2	6	3
3	1	1	1	1	1.1	1	5	2	6	3
4	2	2	2	2	2.2	1	3	4	3	3
5	0.5	1	1	0.5	1.1	1	6	2	6	3
6	2	1	1	2	1.1	1	3	2	6	3
7	2	1.5	1	0.3...	1.15	0.6.	4	3	5	4
8	1	1	3	3	3.1	3	2	2	2	1
9	0.5	2	1	0.25	1.2	0.5	8	4	4	5
10	0.5	1.5	1	0.3...	1.15	0.6.	7	3	5	4
11	1	2	1	0.5	1.2	0.5	6	4	4	5
12	0.5	1.5	1	0.3...	1.15	0.6.	7	3	5	4
13	2	1.5	3	4	3.15	2	1	3	1	2
14	0.5	1.5	1	0.3...	1.15	0.6.	7	3	5	4
15	0.5	1	1	0.5	1.1	1	6	2	6	3
16	1	1	1	1	1.1	1	5	2	6	3
17	0.5	0.5	1	0.3...	1.05	2	7	1	7	2
18	0.5	1.5	1	0.3..	1.15	0.6	7	3	4	4

Table 2 - values and ranking of investigative priorities under different approaches.

Problems with organizationally assigned weights include, without limit, the arbitrary nature of the derivation of w and the potential exploitation of w by the opposing party to assert reliability figures or challenge process credibility. An alternative approach is to assign w based on the number of chains involving the event. By searching for traces of events that cause failures in more evidence chains first, the investigation terminates. with fewer traces examined This assumes that within a chain, all events have equal weight. Sorting can be done based on results, using the number of chains as the primary and c as a secondary criterion, the number of chains over c as the criteria, and so forth.

For the example of [2], the calculations in Table 2 indicate the event number (e), the give weight (w), the given cost (c), the number of chains involved (n), the ROI result (ROI), results of the number of chains followed by the cost approach (Ch), and the number of chains divided by the cost approach (CR). It also includes the ranking, with lower numbers indicating earlier undertaking of tasks under the ROI approach (R), the minimum cost first (R2) approach, the Chains approach (RC) and the Chains ROI approach (RCR). Results show that different assumptions regarding utilities produce different investigative orderings. Depending on specifics of the circumstances and available knowledge, each of these approaches has value for the investigating agency.

Cuts of graphs have also been studied in great depth in the mathematical literature and may be used to identify minimum cost approaches to severing graphs, including the severing of evidential chains such as those proposed in [2]. Cuts are formed by applying the challenges identified in [3] to the graphs of [2]. For example, severing event 13

from [2] reduces the overall weight of the evidence to below half of the maximum weight, making charges appear infeasible both for criminal matters ("beyond a reasonable doubt"), and civil matters, ("the preponderance of the evidence"). If "Internet connection is available at the relevant time" is shown not true, the entire case falls apart. While the absence of evidence does not necessarily imply that evidence of absence, the absence of evidence of availability of access eliminates the potential to prove opportunity in the manner supposed by the hypotheses proposed.

The same optimization criteria may be used by those trying to challenge evidence as those seeking to support it. But challengers to evidence have the added advantage if they can find traces indicative of innocence. For example, if there is a trace of the defendant present at a different location at the time of the incident, this trace may reduce the weight of the event chains to zero. Strategies may differ based on standard of proof, the percentage of successful cases, discovery rules, and tactics. Game theory may be well applied to analysis of this approach, for example to help determine what to disclose when, what to investigate, and risk vs. benefit of tests for different traces.

In [2], a detailing of the structures for each class of cases is required. This may be facilitated by the policy languages such as those in [4] and in widely published papers on policy analysis. Without a definitive framework, a challenger may assert that there are many other possible traces that were ignored and that those traces might refute the claims. The detailing of an approach also begs the question of thoroughness.

For practical purposes, optimization strategy may be experimentally determined. The cost of operating models simultaneously is negligible once one model is automated. A set of models may be run simultaneously with the assignment of cases prorated to existing evidence of success. On a case-type by case-type basis, different optimization may be called for, and this can be facilitated by this strategic approach, as can adaptation to changing conditions over time. However, the fundamental limitations of this model remain regardless of the optimization strategy used. A more basic question is whether and when this model is beneficial. To compare, alternative models are required.

4. A proposed alternative model

We propose an alternative to the model identified in [2] in which detailed trajectories through legal requirements associated with charges are associated with event chains as in [2], supported by traces, which allow traces to be applied to multiple events, and in which costs are associated with each step of the examination of the elements of these chains, but augmented in several ways over [2]. Any such model, in order to be meaningful must sit in the larger context of a physics. Such an information physics for this model is defined in [9], and is beyond the scope of this paper.

The proposed alternative model with an embedded and slightly altered example from a real case using a real law [6] is characterized as follows.

The legal context

A legal statute, or law (L) is associated with a violation (V), consisting of a logic expression $L:\{l_1, ..., l_n\}$, $R:\{r_1, ..., r_m\}$, $LxR \rightarrow [F|T]$, where l_x is an element of the statute and R is a relationship between elements of the statute so that if the set of elements required to meet the relationship defining a violation (the truth of LxR) are present, it implies that a charge of violation is warranted based on the defined legal criteria. ($LxR \Rightarrow V$) For example, [6] a US Federal statute reads, in part, "(a)" Whoever, [for commerce] knowingly ... (3) materially falsifies header information in multiple

commercial electronic mall messages and intentionally initiates the transmission of such messages... shall be punished...". This statute (L) can be broken down into elements including (l_1) the act was for commercial purposes, (l_2) there is material falsification of a header, (l_3) the falsification is present in more than one email message, (l_4) the actor initiated the transmission of these messages, and (l_5) that initiation was the intent of the actor. All of these must be proven to within the standard of proof by the charging party in order for the punishment to be invoked, and the resulting expression might be of the form $L=(l_1*l_2*l_3*l_4*l_5)$.

The hypothesized claims

Claims, (i.e., hypotheses), ($H=\{H_1, ..., H_n\}$) are made in the form of statements which may be supported or refuted by DFE and which support or refute V. For example, (H_1) Defendant sent email messages accompanied by falsified, misrepresented, or forged header information and (H_1) Defendant sent or caused to be sent at least 26,000 false and/or deceptive commercial e-mail advertisements to Plaintiff (P) servers.

The hypothesized events

For each element of the legal requirement [$\forall l \in L$] there is a set of event claims [$E: \{E_1, ..., E_o\}$], each consisting of a set of indicated events from the set of all events [$\forall e, e \in E^*$] within and outside of the digital system [$\forall E_x \in E$, $E_x:(e_{x1} \in E^*, ..., e_{xp} \in E^*)$], and that, in combination, purport to constitute a demonstration that $LxR \Rightarrow V$. Again, from the example, (E_1) Emails received by Plaintiff contained or were accompanied by falsified, misrepresented, and/or forged header information, (E_2), Emails received by Plaintiff had subject lines designed to mislead a recipient regarding the contents or subject matter of a message, etc. And as events, (e_a) The “HELO” protocols on some of the emails provided “identities” of the sending computers that do not match the IP addresses of the sending computers, (e_b) The “identities” provided by Defendant and/or its agents or the computers delivering the emails do not match the IP addresses of the contacting computers, etc. Such claims may or may not be reasonable, logical, or internally consistent, may include assertions made by counsel and facts or statements taken from other sources, including those of the parties involved. The DFE examination issue is to confirm or refute these events.

The traces

There is the set of possible traces from existing evidence $T:(t_1, ...,t_q)$, each element of which may exist. T may be incomplete in that $\exists t: t \notin T$. Subsets ($\tau \subseteq T$) tend to support or refute events relevant to the matter at hand. In the example, only a small subset of the asserted events can be confirmed or refuted by DFE, and in many cases, only select elements of asserted events produce probative traces. For example, there may be classes (c) of traces within emails of many different events, including, without limit, (c_1) date and time stamps of "Received:" headers, (c_2) IP addresses from audit trails, (c_3) date and time indicators in "From " separators, etc. There are many such traces in most cases.

Internal consistency between traces

There is an internal consistency relation $C:TxT \rightarrow [-1...1]$ between traces, and $\forall c \in C$, $c \rightarrow [-1...1]$. All sets of traces relate to all other sets of traces by ranging from completely inconsistent (-1) to completely consistent (1), with 0 indicating that the relationship is not revealing. For example; correspondence of dates and times from email "From " separators and "Received:" headers may be either consistent or not and to different extents, with inconsistency leading to a potentially reasonable claim of spoliation; sequences of times in "From " separators within mailbox files that are not in

time ordering indicate a fabricated mailbox file rather than original writing; mismatches of "From " separators with other parts of the email sequence content are indicative of fabrication or spoliation; sets of different "From " separators with identical email headers and/or bodies, sets of identical "From " separators with differing email headers and/or bodies, and sets of "From " separators with identical dates are either consistent or not consistent with sets of Received:" headers within those same sequences of bytes are all inconsistencies indicative of fabrication or spoliation.

The demonstration consistency of traces

There is another consistency relation $D:TxE^* \rightarrow [-1..1]$, demonstration consistency, that relates all possible traces T and all possible sets of identified events E* and which may tend to confirm or refute hypothesized sets of events by ranging from event sets completely inconsistent with traces (-1) to event sets completely consistent with traces (1), with 0 values indicating that the relation is not revealing. For example; from the class "(c_2) IP addresses from "Received:" headers", if a particular email that is claimed as a violation by Defendant has a trace indicative of an IP address of a competitor of Defendant, an event "(e_h) Plaintiff received commercial email messages sent by Defendant and/or their agents" would seem to be refuted for that email by that trace. Similarly, date and time stamps may be probative with regard to stature of limitations issues. As in [2], there may be many relations between traces and events, so D is many to many onto. The presence of traces does not imply that those traces are reliable. For example, computer dates may be incorrectly set, and emails may be forged. The strength of a refutation depends on the accuracy of the traces. Additional relationships, such as the use of an anchor event in conjunction with another event [7] may result in a higher value of the relation. Thus there is a synergistic effect between elements in subsets of D so that the combination of several traces may cause a far different weight than the sum or the product of the individual weights.

The forensic procedures

There is a finite set of forensic procedures $P:\{p_1, ..., p_n\}$, $\forall p \in P$, $p \rightarrow \mathbb{c} \subset C$, $p \rightarrow \mathbb{D} \subset D$, $p \rightarrow \mathbb{c} \not\subset C$, $p \rightarrow \mathbb{D} \not\subset D$ available to the forensic examiner. Procedures are normally implemented using methods and tools that have properties. Each procedure has the potential to act on any subset of T and to produce false positives (make), false negatives +(miss), or correctly find the presence or absence of subsets of C and/or D. For example, the use of the program "grep" in a Unix-like operating environment may be applied to traces to seek instances of strings typical of IP addresses within areas of traces typically associated with "Received:" headers, and the presence of particular IP addresses identified as belonging to the Defendant may tend to support a particular event. However, the "grep" command may or may not be applied in such a manner as to produce false positives or false negatives, and thus it may make or miss connections between the traces it is applied to and relevant events. While legally, in most jurisdictions, all procedures are theoretically available to all parties, some procedures, either because they are not published, are prohibited, or because examiner are unaware of them, may not be known to or available to any or all parties at any or all times.

Available resources

Each party has finite resources R:(T,$,C,E). Procedures consume time, money, capabilities, and expertise, and each of these elements limit the ability of the parties to fully examine the space of possibilities. In [2], as discussed above, a simple model of cost is used to represent resource constraints, but in general, the resource problem in digital forensics corresponds to the resource problem in other fields of optimization.

The schedule

A schedule sequence $S:(s_1, s_2, ...), \forall s \in S$, $s:(L \subset L, R \subset R, H \subset H, E \subset E, T \subset T, C \subset C, D \subset D, P \subset P, R \subset R, t, t')$ exists where t and t' bound the time period for each step in the schedule, and only subsets of L, R, H, E, T, C, D, P, and R are available within that time frame. Arguments asserting and refuting claims are made to triers of fact (judges or juries) in a sequential fashion with one side presenting then the other, a limited number of "rounds" of presentations are available, specific time frames and similar constraints are placed on all such information exchanges, and the standards of proof, ability of the triers of fact to understand the arguments, and space available for presentation of arguments and facts vary with case type, jurisdiction, triers of fact, and situation. [9] The number of possibilities is clearly large, and the impact on the matter may be profound.

Even though, from a logical standpoint, adequate confirmations or refutations may exist to secure a theoretical confirmation or refutation of the charges, the actual legal matter may have an outcome that is inconsistent with the result of the logical analysis. For these reasons, a single confirmation or refutation is generally considered inadequate and, especially when a great deal is at stake and the participants have adequate resources to do so, more complete exploration of the space is undertaken. All of these impact the ability to and order of the search the space of T and E and the search for relations C and D, and this affects the schedule. In cases where digital forensic issues are important, the potential consequences are high, and adequate resources are made available, a larger portion of the space of {L, R, H, E, T, C, D, P, R} is likely to be explored.

5. Preliminary analysis of the proposed alternative model

Within this model, certain things are clear. While the model of [2] is likely to be useful for making decisions, it represents only a small subset of the issues involved in DFE examination. The smaller model is more practical for the purpose it is being applied to, but it is also less realistic. Depending on the nature of the challenge being met, different subsets of the proposed model may be applied and specific assumptions stated, with those now stated assumptions being made clear by the selection of the subset of model elements. For example, the model of [2] can be seen in the context of the proposed model to ignore all but a single element of S, and for that element, it ignores all but a single subclass of R, assumes consistent E, ignores the details of T, C, and P, and assumes a single metric for D and attends only to $ within R.

The size of the search space involved in any substantial matter is enormous, and thoroughly searching it or even achieving substantial coverage of it, for any nontrivial matter, is infeasible. The present model indicates, specifically:

L is finite, and for any given matter, it is defined by the specific laws.

R is usually expressible as a combinational logic expression, with metric thresholds.

H is unlimited in possible makeup, but in any particular case, H is defined by documents, and courts prevent alteration of H beyond some time within the schedule.

E can be very large, but in most cases it is a few hundred to a few thousand asserted events including statements by the parties in depositions, testimony, and so forth.

T is the size of all sets of all states of all digital automata in existence at all relevant times. But in any particular matter, T is limited to the available traces. This reduced T, however; is very large. Every possible subset of available bits can constitute a trace. For

8 bits, there are 2^8 different sets of bits that can comprise traces, and for each of those sets of bits, there are 2^n different possible traces, where n is the number of bits in the trace. Generally, there are m!n sets of bits of length n in a collection of m bits, and for each of those, there are 2^n different possible traces. The total number of traces for m bits of data is then $\sum(m!n)2^n$ for n=1 to m. So the set of all possible traces for a single byte comes to $\sum(8!n)2^n$ for n=1 to 8, or 6560 unique 8-bit traces. For 16 bits, this comes to 43046720, and for 64 bits, it comes to more than $3*10^{31}$. Clearly, for any substantial set of bits, the space of traces cannot be exhausted. The evidence identification problem is fundamentally about identifying relevant subsets of T, and this problem is not even close to being solved. However, legal precedent in the United States has led to the requirement to preserve evidence that might reasonably be believed to be relevant to the matter at hand, as of the time that any party has or reasonably should have knowledge that the evidence may be material. Thus the parties have an obligation to diligently identify and preserve or cause to be preserved, traces like audit trails from contractors and providers, content from related systems, and any other such traces.

C is $|T|^2$. For substantial T, this is very large. For 64 bits of total evidence, the size of the set of all internal trace consistencies and inconsistencies is approximately 10^{63}. This makes any notion of coverage of C by exhaustion infeasible. It may appear that many traces are independent of each other, but there may be subtle interactions. For example, a time stamp of user data entry to a database on one computer may relate to a trace of a Web page lookup on a seemingly unrelated computer through the deviation of timing on the data entry caused by a domain name system (DNS) lookup delayed due to the DNS lookup of the Web request. While finding and associating such a trace may seem nearly impossible, it seems clear that in the interconnected world of the Internet, subtle effects exist and may leave traces. In practice, a relatively small set of traces may be examined, and the selection of the traces to be examined and method for doing such examination is not well defined or developed except in specific areas.

D is |T|*|power set of E|. That is, each subset of traces may interact with each subset of claimed events. This is again too large for practical exhaustion. As for C, the re may be subtle interactions between distant traces and asserted events, and subtle effects may produce relations that are hard to identify. This goes to the problem of trace identification and collection as well as analysis. As with C, a relatively small set of traces may be examined for a subset of the event sets, and the selection of the traces to be examined for event sets and method for doing such examination is not well defined or developed except in specific areas.

P is the size of all possible instruction sequences executed on all subsets of T and E in the context of all possible initial memory states over a defined time. The number of possible instruction sequences of a given length is on the order of the number of different instructions in the processor taken to the power of the number of instructions that can be executed in the defined time. For 100 instructions and 10^9 instructions per second, the number of instruction sequences comes to a number written as a 1 followed by 10^{18} 0's. This its then multiplied by the number of possible initial memory states and by the size of D to get the number of possible analyses that can be done in one second of computer time. Clearly the number of procedures it too large to seriously contemplate and any actual procedures executed will cover only a very small subset of the possible procedures. In practice, the number of actual procedures available is very small, limited to the number of procedures developed by people or their machines. The number of procedures that meet the legal requirements of being scientific according to a

defined methodology properly applied, being executed by tools that have been tested, calibrated, demonstrated to be reliable, and properly apply the defined methodology is far smaller still. There are, perhaps, a few thousand such procedures in digital forensics today, and the number that have been published and peer reviewed is smaller still.

R constrains process. Time constraints are forced S, timings of actions of parties, the complexity of procedures, the expertise of the examiners, available computing capacity, and costs. Expertise is typically a limiting factor because it involves people with knowledge, skills, training, experience, and education that can combine understanding of the legal situation, understanding of technology, and programming and operations, to create analytical methods that both meet the needs of the legal system and are revealing with respect to the matter at hand.

S acts to constrain the process in real-time and alters the nature of the digital forensic effort over time, sometimes quite dramatically. Depending on the specifics of the legal matter, the total time frame from first notice of a legal matter to final disposition may be as short as a few weeks or as long as tens of years. Typical matters are resolved in less than two years, and deadlines are commonly on the order of weeks apart.

6. Summary, conclusions, and future work

A new model was introduced that depicts the inherent nature of digital forensic trace evidence in the legal context, and the result is that many problematic issues are identified. These include, without limit: (1) We don't have a theoretical framework for identifying all of the meaningful traces and their relationships. (2) The size of the space of traces implies that though examination is almost always likely to be infeasible. (3) Synergistic relationships exist between different elements of T and E so that basic properties such as independence and transitivity do not necessarily apply. (4) There is no uniform framework for evaluation of strategies and the number of possible strategies for trace analysis is enormous. (5) Reliability figures and error rates associated with procedures are essentially non-existent in most procedures today and no widely accepted fault models exist today for gaining insight into reliability of such procedures. (6) Measures of coverage, notions of adequacy of trace analysis, and the relations between traces and events are unable to meaningfully address coverage on their own.

References

[1] Gregory H. Carlton and Reginald Worthley, "An evaluation of agreement and conflict among computer forensics experts", Proceedings of the 42nd Hawaii International Conference on System Sciences - 2009.

[2] R. Overill, M. Kwan, K. Chow, P. Lai, and F. Law, "A Cost-Effective Forensic Investigation Model", IFIP WG 11.9, International Conference on Digital Forensics, Jan 25-27, 2009.

[3] F. Cohen, "Challenges to Digital Forensic Evidence", ASP Press, 2008 ISBN#1-878109-41-3

[4] F. Cohen, "Policy-Based Security and Enterprise Policy Management", Burton Group, Security and Risk Management Strategies Report, 2003-12-09

[5] K. Inman and N. Rudin, "Principles and practices of criminalistics: the profession of forensic science", ISBN# 0-8493-9127-4, CRC Press, 2001

[6] 15 USC 103 ‘‘Controlling the Assault of Non-Solicited Pornography and Marketing Act of 2003’’, or the ‘‘CAN-SPAM Act of 2003’’, http://uscode.house.gov/download/pls/15C103.txt

[7] F. Cohen, "Issues and a case study in bulk email forensics", Fifth Annual IFIP WG 11.9 International Conference on Digital Forensics, Orlando, Florida, USA, January 25 - 28, 2009, also appearing in "Advances in Digital Forensics V".

[8] M Kwan, K P Chow, F Law & P Lai, Reasoning About Evidence Using Bayesian Networks, Advances in Digital Forensics IV, 2008, pp.141-155.

[9] F. Cohen, "Digital Forensic Evidence Examination", ASP Press, 2009, ISBN#1-878109-44-8. This book contains a more complete version of this paper as a chapter.