Two Models of Digital Forensic Examination
May 21, 2009

Dr. Fred Cohen
President - California Sciences Institute
CEO – Fred Cohen & Associates
Outline

- Background and Introduction
- An existing model
- Analysis of the existing model
- A proposed alternative model
- Analysis of the alternative model
- Summary, conclusions, and further work
My background

- **California Sciences Institute**
  - 501(c)(3) non-profit California research and educational institution - WASC accreditation candidacy pending
  - Ph.D. Program in digital forensics (Fall 2009)

- **Fred Cohen & Associates**
  - Enterprise information protection consulting
  - Digital forensics (high fees – no guarantees)

- **Fred Cohen – Digital forensics**
  - POST certified instructor, FLETC instructor, books and book chapters, papers, testimony in Federal, State, and Local courts
Previous models

- Carrier and Gladyshev
  - Model the forensic analysis process in terms of consistency and inconsistency and introduce various time-related concepts

- Stallard and Levitt
  - Semantic integrity checking (consistency)

- My basic notion and approach
  - If we are going to make a science of digital forensics, we need to develop a physics and a theory for applying that physics
  - This paper is about a theoretical model
Basic notions of forensics

• The evidence is a set of traces
  – A “trace” is a “bag of bits”
  – Normally an ordered sequence
  – It is the result of some digital process
  – The question is: “What process?”
  – How do we find out?
  – How sure are we? Why are we this sure?

• The evidence is latent in nature and technical
  – You need tools to see it and experts to explain it
  – What tools, and how can you trust them?
  – What experts, and how credible are they?
Outline

- Background and Introduction
- An existing model
- Analysis of the existing model
- A proposed alternative model
- Analysis of the alternative model
- Summary, conclusions, and further work
• A model of making decisions
  – About processing evidence in cases
  – Prioritizing resources based on likely outcomes
  – Modeling the legal process with the evidence

• The basic model
  – A legal requirement for a violation \( L: \{l_1, \ldots, l_n\} \rightarrow V \)
  – Sets of evidence chains \( E: \{E_1, \ldots, E_o\} \) show \( L \)
  – Traces demonstrate evidence \( T: \{t_1, \ldots, t_n\} \rightarrow E \)
  – Evidence has weights and they sum
  – Enough weight and you exceed the \( V \) threshold
How a case is made

- Previous cases provide precedent
  - Necessary evidence chains to get a conviction
- Investigation takes resources
  - Desire to minimize resources per conviction
- Figure out how to spend resources
  - Identify $T \rightarrow E \rightarrow V$ and costs for each $t \in T$
  - Order investigation to find $t \in T$ for minimum cost
  - Go one step through $E$ at a time
  - Since refutation cuts $E$, stop when $E$ is cut
  - If cost effective, try alternative $E$s
Outline

• Background and Introduction
• An existing model
• Analysis of the existing model
• A proposed alternative model
• Analysis of the alternative model
• Summary, conclusions, and further work
Kwan's optimization approach

- Problems include, without limit:
  - E is a POset
  - No method for evaluating costs or thresholds
  - Cost of a node in the POset has rewards for all Posets passing through the node
  - If a node is refuted, it cuts all Posets passing through it
  - Different valuation models produce different ordering of nodes for optimization
  - The method being used potentially leads to gaming of the system for the criminals
  - Clever criminals can optimize their activities to defeat prosecution (others get caught first)
Outline

- Background and Introduction
- An existing model
- Analysis of the existing model
- A proposed alternative model
- Analysis of the alternative model
- Summary, conclusions, and further work
California Sciences Institute is a 501(c)3 non-profit educational and research institution. We do not discriminate in our hiring, admissions, offerings, or in any other way except by ability to do the work and learn the material.

Fred Cohen & Associates

California Sciences Institute is a 501(c)3 non-profit educational and research institution. We do not discriminate in our hiring, admissions, offerings, or in any other way except by ability to do the work and learn the material.
The new model

- Laws: $L: \{l_1, ..., l_n\}$, $R: \{r_1, ..., r_m\}$, $L \times R \rightarrow [F|T]$
- Violations: $V: L \times R \rightarrow [-1 ... 0 ... 1]$
- Hypothesized claims: $H=\{H_1, ..., H_n\}$, $H \subset V$
- Events: $E: \{e_1, ..., e_o\}$
  - Filings, statements, etc. non DFE
- Traces: $T: (t_1, ..., t_q)$ \{all subsequences of $T$\}
  - All subsets of the bag of bits
- Trace (internal) consistency: $C: T \times T \rightarrow [-1...1]$
- Demonstration consistency: $D: T \times E^* \rightarrow [-1..1]$
New model (continued)

- $P:\{p_1, \ldots, p_n\}, \forall p \in P, p \to \{c \subset C, d \subset D, c \not\subset C, d \not\subset D\}$
  - The forensic procedures confirm or refute type C and type D consistency

- Resources $R: (T, $, C, E)$
  - Time, Money, Capabilities, and Expertise

- The Schedule $S: (s_1, s_2, \ldots), \forall s \in S$
  - $s: (t \subseteq L, r \subseteq R, h \subseteq H, e \subseteq E, t \subseteq T, c \subseteq C, d \subseteq D, p \subseteq P, r \subseteq R, t, t')$
    - The schedule is a sequence of spans of time in which laws, relations, hypotheses, events, traces, type C and D consistency and inconsistency, forensic procedures, and resources apply.
An email header

- Original writing
- Received in New Jersey

Type C

Type D

Example: an email extract

From: ??@?? Fri, 15 May 2009 02:39:41
Return-path: <svein@willassen.no>
Received: from smtpin126-bge351000 ([10.150.68.126])
  by ms283.mac.com (Sun Java(tm) System Messaging Server 6.3-7.04 (built Sep 26 2009 64bit)) with ESMTP id <0KJP00J852A8S8S10@ms283.mac.com> for
dr.cohen@mac.com; Fri, 15 May 2009 09:39:41 -0700 (PDT)
Original-recipient: rfc822;dr.cohen@mac.com
Received: from mail-bw0-f162.google.com ([209.85.218.162])
  by smtpin126.mac.com (Sun Java(tm) System Messaging Server 6.3-8.01 (built Dec 16 2008; 32bit)) with ESMTP id <0KJP0018P29JIHD0@smtpin126.mac.com> for
dr.cohen@mac.com (ORCPT dr.cohen@mac.com); Fri,
15 May 2009 09:39:41 -0700 (PDT)
X-Brightmail-Tracker: AAAAAA==
Received: by mail-bw0-f162.google.com with SMTP id 6so3067145bwz.30 for
  <dr.cohen@mac.com> Fri, 15 May 2009 18:39:41 -0400 (EDT)
MIME-version: 1.0
Received: by 10.204.57.138 with SMTP id c10mr3481822bkh.56.1242405581619; Fri,
In-reply-to: <C93BF973-C2E2-4CA7-B77B-EB48283A4028@mac.com>
Date: Fri, 15 May 2009 18:39:41 +0200
Message-id: <2e67f5b00905150939r2e34c9d9n96688c4ac7f5ea98@mail.gmail.com>
Subject: Re: A question on your dissertation and an experiment to try
From: Svein Yngvar Willassen <svein@willassen.no>
To: Cohen Fred <dr.cohen@mac.com>
Content-type: text/plain; charset=UTF-8
Content-transfer-encoding: quoted-printable

Fred Cohen & Associates California Sciences Institute is a 501(c)3 non-profit educational and research institution. We do not discriminate in our hiring, admissions, offerings, or in any other way except by ability to do the work and learn the material.
What's the problem?

- Type C problems identified (so far)
  - “From “ separator @??? and date format
  - “From “ offset from last Received (False+)
  - Received: times in the same second (how fast?)
  - Gmail message-ID but emitted from non-gmail account (passes through Google later – Google added AFTER earlier “Received:”?)
  - Message server built after Message Received!
  - Server versions inverted w.r.t. Build time stamps
- Type D problems identified (so far)
  - Received in NJ inconsistent with all time zones
- Lots of traces extracted from the original trace
California Sciences Institute
This is only the beginning

- Which if these are actually spoliation?
  - And how do we tell?

- How many more traces are there?
  - In this specific sequence?
  - Are there other sequences?
  - How about cross-sequence C consistency?

- How do these relate to other events?
  - Version numbers of servers and dates and times
  - Anchor events tying down other facets
  - Character sets available on machines at times

- Where does it end?
Outline

- Background and Introduction
- An existing model
- Analysis of the existing model
- A proposed alternative model
- Analysis of the alternative model
- Summary, conclusions, and further work
The size of the space

- L is finite, and defined by the specific laws.
- R is usually expressible as a combinational logic expression, with metric thresholds.
- H is unlimited in possible makeup, but H is defined by documents, not very alterable and time limited by 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.
More sizes

- $T$ is the size of all sets of all states
- In a particular matter, $T$ is the available traces
- For $m$ bits of traces, $|T| = \sum (m!n)2^n$ for $n=1$ to $m$
  - 64 bit trace $\rightarrow 3 \times 10^{31}$ possible actual traces
- $C$ is $|T|^2$
  - 64 bit trace $\rightarrow 10^{63}$
- $D$ is $|T|^\ast |\text{power set of } E|$

Exhausting $C$ or $D$ is infeasible for any real case

- Exhausting consistency checks is infeasible
- What is a “thorough” job?
Forensic procedures

- $P$ is the size of all instruction sequences executed on all subsets of $T$ and $E$

- $|\text{Instruction set}|_{\text{number of instructions executed}}$
  - 100 instruction instruction set
  - $10^9$ instructions per second for 1 second
  - $|P| \approx 1$ followed by $10^{18}$ 0's.

- $|P|$ in reality is – perhaps $10^3$-$10^4$?
  - scientific methodology properly applied
  - executed by tools that have been tested, calibrated, demonstrated to be reliable
  - Applied by suitable experts
Resources and schedule

- R and S constrain process
  - Time limits → limited P and exploration of C/D
  - Money limits → limited P, time, capabilities, expertise
  - Capabilities limit → limited P
  - Expertise limits → limited P

- S changes with time and situation
  - The sands literally shift underneath you
  - No analytical methods are available to optimize at this level of complexity
  - Game theory doesn't come close to it
  - The skill of the participants rules the day
Returning to the example

- How many more traces are there?
  - We now know the answer – and it hurts!

- How many more procedures may there be?
  - An enormous number in total – but which are probative and how reliable are they?
  - We don't even know how many more there may be for a single email header!

- How do we test the reliability of the apparent inconsistencies?
  - We need an experimental base and samples and lots of procedures to test
More on the example

• Resources are constrained – even for this email

− How do we find out about the Message-ID field in context of other similar fields?
− How do we identify the source of the version number/time inversion problem?
− We haven't even looked up the IP addresses vs. host names and time zones
− What about the internal ESMTP IDs? Are they in proper sequence?
− Is Google really adding GMAIL Message-IDs to all non ID'd messages?
− Is the originator on a 10-net using the proper ...
Outline

- Background and Introduction
- An existing model
- Analysis of the existing model
- A proposed alternative model
- Analysis of the alternative model
- Summary, conclusions, and further work
Summary

• Earlier models are less comprehensive
  – The new model is more so
  – Optimization in previous models was problematic – but this one is no better

• The present model
  – Clearly shows complexity challenges with traces and examination of traces
  – Shows the size of the problem space for what it is and dispels any notions of “comprehensive”
  – Brings a notion of how to apply redundancy to understanding trace and event consistency
  – Introduces type C and D consistency
Summary

- Procedures are extremely limited today
  - Major effort is needed to create and test new procedures for types C and D consistency
  - Understanding the class of P seems important

- Resource limits and schedule
  - The notion of resource limits and schedule introduce a more complex and more realistic optimization arena
  - Many new challenges appear to be put forth by this model and its potential application
  - Game theory appears to be too weak for this class of problems – at least as it exists today
Conclusions

• We have the start of a scientific methodology
  – We now know that being “comprehensive” or “thorough” in examination of DFE is infeasible
  – We now know why this is so, and why it will likely remain infeasible for quite some time
  – We now have a theoretical model for developing metrics associated with examination
  – We have a basis for identifying complexity issues with forensic procedures
  – We can use the model along with complexity analysis to allocate resources within schedules

• But it's only a start
Future work

• A model is only a model
  – The development of the science of DFE examination is in its infancy
  – We need a well defined and accepted physics
  – We need to develop systematic and scientific procedures for type C and D consistency
  – We need clarity around the methodology and its proper application
  – We need to start to do complexity analysis to understand what is and is not feasible

• But without a model, we grope in the dark
http://calsci.org/ - calsci at calsci.org
http://all.net/ - fc at all.net
Further Reading