Digital Forensics Research: The Good, the Bad, and the Unaddressed

by Nicole L. Beebe, Ph.D.
5th Annual IFIP WG 11.9
January 27, 2009
Discussion Topics

- General Successes
- General Failures
- Research Needs
Background (Bias)

- My background
  - Ex-Law Enforcement (AFOSI, 1998-2007)
  - Private Sector Forensics (2001-Present)
  - Educational background
    - B.S. Electrical engineering
    - M.S. Criminal Justice
    - Ph.D. Information systems
  - Assistant Professor in Information Systems Dept.
Background – Contributors

- Dr. Sujeet Shenoi – Univ. of Tulsa
- Mark Pollicit – Ex-FBI, Univ. of Central Florida
- Eoghan Casey – Stroz Friedberg LLC, Johns Hopkins Univ.
- Dr. Simson Garfinkel – Naval Post Grad School (Harvard, MIT)
- Eric Thompson – CEO, Access Data Inc.
- Ovie Carroll – Ex-AFOSI, DoJ CCIPS Cybercrime Lab Dir.
- Dave Baker – MITRE
- John Garris – Ex-AFOSI, NASA OIG Computer Crimes SAI/C
- Randy Stone – Detective, Wichita Police Dept.
- Dr. Marc Rogers – Purdue Univ.
- Dr. Frank Adelstein – ATC-NY
- Dr. Wietse Venema – IBM
Background – Contributors

- Gary King – AFOSI Computer Crime Investigations Program Mgr
- Dr. Florian Ducholz – James Madison Univ.
- Dr. Vassil Roussev – Univ. of New Orleans
- Jesse Kornblum – ManTech
- Russell McWhorter – Bexar County Sheriff’s Office, Veridicus Inc.
- DeWayne Duff – Ex-AFOSI, Stroz Friedberg LLC
- Rod Gregg – Ex-FBI, Stroz Friedberg LLC
- Drew Fahey – Ex-AFOSI, e-fense Inc. (developer of Helix)
- … plus seven other researchers & practitioners
- … and, of course, me.
The Good

- Unequivocal improvement in prominence & value of digital evidence in investigations
- Becoming more scientific
  - Formalization/standardization of processes/approaches
  - Formulating DF problems into scientific research Q’s
  - DF research starting to enter mainstream research
- Archeology of digital artifacts (Windows/Linux)
- Cross-discipline knowledge sharing
- Tackling the DF problem de jour (e.g. memory)

Archeology – reverse engineering, documenting, exploiting data structures (windows/linux OS, FS, applications – content, metadata)
   -- counterpoint – will see later, view that archeological site locations need to be expanded to other OS/FS

Cross-discipline knowledge sharing – CS, IS, CJ, LE, engineering(?)
   -- counterpoint – one missing (info science)

Tackling the DF problem de jour – when CoP/research IDs key research problem/question (data carving, memory acquisition/analysis), community systematically tackles the problem, resulting in more and higher quality research
   -- counterpoint – tackle single problems at the exclusion of all else?
More Kudos

- HW write-blocking industry
- Acquisition/collection phase in general
- Live forensics
- Contributions to qualification, certification, etc.
  discussion
- Honorable mentions
  - The Sleuth Kit (high quality, open-source tool)
  - AFF (vendor neutral, compress-able imaging format)

HW write-blocking industry – really a research community contribution? Certainly not academic research community contribution.
The Bad

- Hyper-formalization of processes/approaches
  - Agencies getting dangerously close to checklists
- Cross-discipline knowledge sharing incomplete
  - Lack of extension of information science research
- Insufficient research into other OS & FS
  - HFS+, UFS, ZFS, proprietary systems, etc.
- Data-centric not info/knowledge-centric
  - Researchers & practitioners are both guilty
More Criticisms

- Lacking a common body of knowledge
- Accreditation “machine” has spun out of control
- Bridging the gap between research & application
- Still lacking rigor & relevance in research
- Lack of a clear research agenda
  - Common CFP topics, but cover full-spectrum
  - Lack of federal funding of research (U.S. complaint)
  - Commercial industry shaping the agenda
    - Decisively toward e-discovery research questions

Common body of knowledge
- Need for knowledge/skills assessments
- Need for quality in educational programs
- Need for extending non-digital concepts (legal, accreditation, etc.) to the digital domain

The ‘accreditation machine’ has grown out of control – started to meet a legitimate need for quality assurance/control, but has grown to a degree of complexity because of the industry it created and the industry’s lack of understanding of the discipline that it now results in 40% overhead on case processing time/resources. “Too often, we have found ourselves so involved in the process that we’ve lost sight of the purpose.”

Bridging gap between research & application
- Common complaint by BOTH researcher & practitioners that we’re not producing usable tools from our research
- Is tool development a research activity???
- If it is, then our tools must be developed using better s/w engineering principles (cuz not being adopted now)

Rigor & relevance
- still lots of junk research out there
- increased publication venues/interest cause junk to be published
- lots of duplicative research; ‘wheel reinventing’
- standard for knowledge contribution is barely even incremental
The Unaddressed

(or at least needing more attention)
Volume & Scalability

- Acquire & process more faster
- Logical acquisitions – decision support systems
  - And/or non-“complete” physical acquisitions
- Collaborative, distributed analysis
  - Collaboration management
  - Data storage/transfer (centralized, decentralized)
- Lagging S/W development
  - H/W advances (multi threading / massive parallelism)
  - Tools to handle large volumes of email
- Data analytics, linkages & pattern analysis

Won’t work: Traditionally & continually view it as a classical database problem
- index & query

Won’t work: Throwing more hardware at the problem won’t solve the scale issue

Collaborative, distributed analysis
DF data warehousing, data mining; centralized/decentralized data management
Store images centrally to facilitate data mining, but ability to analyze
decentrally, maybe even collaborative analysis in disparate locations
Need tools that can handle partial image data transfer, not full image files
~“centralized repositories, but manage distributed
access/searching/correlation”

Email: Need tools that can analyze large volumes of email w/o introducing error
and/or don’t crash!
Intelligent Analytical Approaches

- Need to extend artificial intelligence and other intelligent search/retrieval algorithms/approaches
  - Semantic vs. literal searching techniques
  - Improved data indexing / relational data optimization
- Similarity matching mechanisms
  - “Fuzzy hashing” requires paradigm shift & scientific certainty research/support
- Intelligent password recovery
  - Passphrase ID/extraction
  - Probabilistic approaches (length, location, signatures)
  - PW caching moving to CPU cache
Small device forensics:
General belief is that every device is forensically different & new devices are coming out before we can reverse engineer them
Counterpoint – belief that industry will eventually move toward more standardization
Counter-counter point – interoperability is currently facilitated via communication protocols, so no need for interoperability of data structures, OS/FS, etc.
“Ease of Use”

- Tool – need to simplify ease of use for practitioner
  - Not too technical
  - Easy to use user interface
  - Protections against human error
    - but allow advanced mode for customizations
- Information – reported findings must be usable
  - Data visualization
  - Cross-correlation, link-analysis (automated)
  - Reduce problem of info overload (need “zoom” capability)
- Paradigm shift from hierarchical to temporal view

PARADIGM SHIFT from hierarchical, relational file/folder view, to graphical timeline/event view (but retain typical analytical capability re: copying, exporting, bookmarking, etc.)
S/W Development/Engineering

- S/W must fully leverage H/W advances
- Increased automation
- Increased interoperability
  - Standardized, interoperable data formats (I/O)
  - Standardized APIs
- Need OS independent DF platforms (e.g. Pyflag)
- Need DF platforms that are all-in-one wrt data
  - Static media, volatile data, network dumps, etc.
“Other”

- Database forensics
- Steganography
- Live file systems
- More work needed on volatile memory analysis
  - Knowledge of disturbance/distortion caused
- Non-windows/linux file systems (HFS+, UFS, ZFS)
- Solid state memory acquisition & analysis
- Investigations involving multiple, distributed systems
- New XML office document standards

Memory analysis: the continuing challenge of conducting live analysis without causing and/or minimizing/measuring/knowing resultant “distortion” caused
Issues of Science

• A way to specify error rates like in traditional forensic sciences
  – Is this realistic?
  – Paradigm shift toward determining/quantifying certainty/confidence?

• Formalization of hypotheses generation & testing
• Repeatable experimentation & comparative eval.
  – Need for a common test corpora
Philosophical Questions

- Is DF field losing its “purity” to e-discovery field?
- Are we immune to DMCA suits?
- Can the research community influence technical specification documentation respecting DF needs?
- Are we keeping pace with anti-forensics research?
Research Agenda Summation

- Volume
  - Non-device level acquisition
  - Intelligent searching, extraction & analysis
- Technological changes
  - Move away from incremental knowledge contributions toward tougher challenges of significant contribution
- Paradigm shifts
  - Non-binomial conclusions (scientifically derived)
  - Conclusion certainty vs. tool/process error rates
- Need to study ease of use, HCI & adoption issues

Non-device level acquisition
-DSS for logical acquisition or non-complete physical acquisition

Intelligent searching...

-Data analytics, linkages & pattern analysis
- Semantic vs. literal
- Extend info science, AI, etc. (IR research) to DF domain

Technological changes (we need to get out of, or better yet GROW our comfort zones)
- Other devices than HDs & removable media
- Virtual & cloud environments
- Other PW extraction mechanisms
- Other OS/FS/Apps

Paradigm shifts
- Similarity matching vs. identical determination

S/W development/engineering – really a DF research question/stream, or a traditional CS one?
Additional thoughts / counterpoints welcome / REQUESTED!