Digital Forensics: Defining a Research Agenda

Kara Nance
Department of Computer Science
University of Alaska Fairbanks
ffkln@uaf.edu

Brian Hay
Department of Computer Science
University of Alaska Fairbanks
brian.hay@uaf.edu

Matt Bishop
Department of Computer Science
University of California, Davis
bishop@cs.ucdavis.edu

Abstract

While many fields have well-defined research agendas, evolution of the field of digital forensics has been largely driven by practitioners in the field. As a result, the majority of the tools and practice have been developed in response to a diverse set of specific threats or scenarios, rather than as the result of a research and development plan. In June, 2008 a group of digital forensics researchers, educators and practitioners met as a working group at the Colloquium for Information Systems Security Education (CISSE 2008) to brainstorm ideas for the development of a research, education, and outreach agenda for Digital Forensics. This paper outlines some of the ideas generated and new research categories and areas identified at this meeting, as well as a plan for future development of a formalized research agenda.

1. Introduction

The idea for this workshop came about during the presentation of the research and education agenda for virtualization and digital forensics. While that particular subcategory was well-defined, it was noted that a comprehensive research agenda for digital forensics was not available and as such researchers, especially Ph.D. students, were finding it challenging to identify topics in this rich research environment. As such the research agenda for virtualization and digital forensics served as a driver and potential model for this more challenging undertaking. The goals associated with defining a research agenda is to provide academic researchers, with challenging and interesting problems in digital forensics and to develop communities of researchers that can work together to advance the state-of-the-art in digital forensics.

2. Background

Unlike many research areas, digital forensics is a largely practitioner-driven field. Advances in the field tend to be primarily developed and applied in reaction to a specific incident or class of incidents. This "bottom-up" approach to digital forensics has made it challenging to identify typologies and develop research taxonomy for this field. Contributing to this challenge is the wide range of fields that have independently contributed to the evolution of digital forensics.

In June, 2008 a group of digital forensics researchers, educators and practitioners met as a Digital Forensics Working Group at CISSE 2008 with the goal of collecting ideas for research categories, research topics, and research problems in digital forensics. The identification of some of the current institutions, organization, and individuals conducting research in specialized categories was a secondary goal. The participants of the Digital Forensics Working Group represented a variety of backgrounds and specializations, with unique perspectives regarding issues within digital forensics. The first task was to try to identify areas of interest, explore categories, subcategories, and provide concrete examples of problems within individual categories.

The long term objective is to distill the identified concepts into a finite number of research agenda items and to describe technical and operational concepts and approaches associated with each identified issue. Finally, a general definition, example, advantages and potential limitations associated with each identified research agenda item will be enumerated in an attempt to formalize an initial research agenda.

While the long-term objective is still under development, the progress made at the initial meetings marks a substantial contribution towards the development of a research agenda for digital forensics.

978-0-7695-3450-3/09 $25.00 © 2009 IEEE

3. Process

The initial meeting of the participants was a free-format brainstorming session in which each individual was given color-coded cards and asked to identify categories, subcategories, and specific research problems in digital forensics. As cards were submitted, they were posted in the front of the room categorically for the entire group to review. As categories were identified, and subcategories and thematic topics associated, the process, in turn, stimulated identification of additional categories and subcategories, as well as reorganization of topics. The process of working together in this manner to identify and associate the research agenda allowed rapid compilation of ideas and themes. Participants were also encouraged to suggest names of organizations and individuals conducting research in each identified area that could be approached for additional information.

4. Findings

Following the brainstorming session, a free-format discussion ensued about the categories, subcategories, and research problems identified during the brainstorming session. While the commentary addressed most of the issues presented, several interesting predominating themes prevailed:

1. Process Control Systems (SCADA Systems) and the lack of associated forensics, legal issues, security, development, education, etc.
2. The challenges associated with educating the diverse constituencies who need digital forensics education and training.
3. The overarching legal issues, both domestic and international, associated with digital forensics.
4. The need to improve the digital forensics collection and analysis processes through parallelization.

As the discussions evolved, the group determined that it might be more palatable to separate legal and educational from the other research categories when constructing a hierarchy as they seems to be overarching themes that could be applied to every research area. It was also determined that the methods for organizing these overarching categories might be significantly different, potentially organized based on the target audience that is being addressed rather than the conceptual content areas.

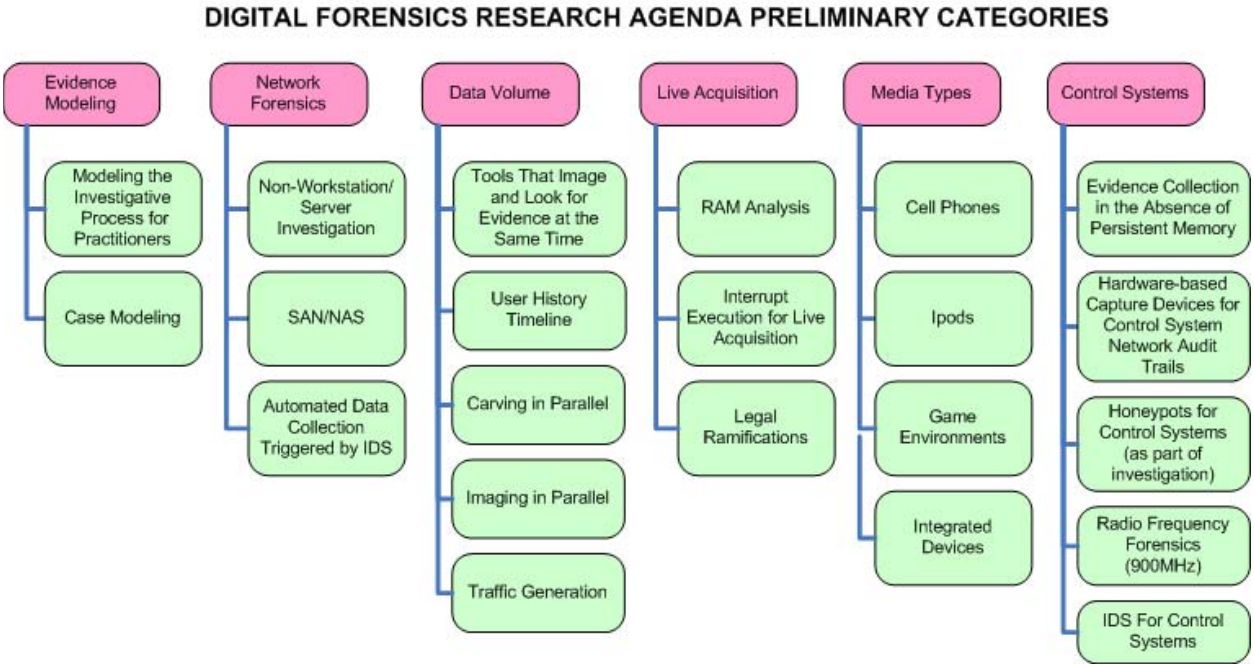

Figure 1: Initial Organization

4.1. Research Agenda

Once the education and legal categories had been isolated from the other research categories, the resulting content areas were more amenable to a hierarchical organization. Beyond Education and Legal, six additional research areas were identified: Evidence Modeling, Network Forensics, Control System Forensics, Parallelizing Data Collection and Analysis, Live Acquisitions, and Media shown in Figure 1.

4.1.1. Evidence Modeling. This category includes modeling the investigative process for practitioners and case modeling for classes of crimes. The objective in this category is to simplify the process for investigators by providing models of evidence that would be associated with particular types of crimes, (e.g., embezzlement, child pornography, etc.). If a model of evidence that would be useful for a particular type of crime is available, it could facilitate all phases of the digital forensics process. These models could be used to guide the investigator in the data collection phase by identifying relevant or potentially relevant data as well as the proper techniques associated with collecting the data. They would also provide a model of data organization for better processing, analysis, and presentation of the data. Finally it would facilitate the presentation phase by providing a pathway for an expert to recall information about a case that they worked on in the past; which can be challenging as the time between the incident and the trial can be substantial.

4.1.2. Network Forensics. While workstations and server network forensics are somewhat well understood, the working group agreed that it is less clear how network data from non end-points, such as switches and routers, can be collected, analyzed, and presented as evidence. In addition, an increasing number of devices that are not workstations or servers now commonly appear on networks. Examples of these non-traditional network devices include office infrastructure (e.g., printers, copiers, scanners, fax machines), media players (e.g., the Apple TV device, the Microsoft Zune, and the Amazon Kindle all have some level of networking ability), game consoles, phones (e.g., cell phones with wireless network capability, or VoIP phones), and even cars (e.g., OnStar connection to cellular networks, or Microsoft Automotive) and home appliances (e.g., an Internet connected display on a fridge). SCADA systems, which are covered in more depth in section 4.1.6, are also increasingly commonly being attached to networks, and typically offer no persistent storage for logging of network activity. The challenge for the DF research community is to develop methods to allow an investigator to determine how these devices interacted with the network during a time period of interest.

The other areas of network forensics that the working group identified as research priorities were associated with automated data collection triggered by an Intrusion Detection System (IDS). This could include incorporating artificial intelligence into an IDS so that the data being collected is based on the "alert level" triggered by the activity being observed.

4.1.3. Data Volume. It is common for digital forensic investigations to be overwhelmed with massive volumes of data. Increasing numbers of devices hold potentially relevant information, and the data storage capacity on such devices is expanding rapidly. It is easy to find examples of digital media players with 160GB hard drives, inexpensive digital cameras that can store 8GB or more, cell phones that have 16GB of flash storage, inexpensive 8GB USB memory sticks, and consumer-grade terabyte hard disks costing no more than a few hundred dollars. In addition, dedicated storage devices offer almost limitless storage volumes, and while Storage Area Network (SAN) devices still tend to be limited to larger corporate environments, consumer-level Network Attached Storage (NAS) devices are available at prices that make them practical for home and small office environments. All of this means that a typical investigation can involve massive volumes of data. While some effort has been made towards parallelization of data processing (such as Access Data's Distributed Network Attack product that parallelizes password recovery across multiple workstations), much more remains to be done if useful information is to be retrieved from these increasingly common large data collections.

Areas identified by the group in which parallelization research could provide benefits included traffic generation, the imaging and carving processes, and the development of user history timelines, including those based on multiple data sources. In addition, approaches that combine data imaging and evidence identification in parallel could also be beneficial, allowing an investigator to potentially direct the data acquisition process based on real-time results to acquire the most promising data sources during the initial phase of analysis.

4.1.4. Live Acquisition. Forensic analysis methods for quiescent systems are currently the norm, and generally involve the acquisition and analysis of the storage media present in the target system. However, these methods generally have no access to the run-time state of the systems, and as a result there can be important information that is not part of the analysis, such as network connections, encryption keys, decrypted data, process lists, and modified code running in memory. As a result there is considerable interest in methods capable of performing analysis on non-quiescent systems, whether this can be done while the system continues to operate normally, or by somehow temporarily interrupting execution while preserving the system state. Research topics identified by the working group included RAM analysis, methods for interrupting the execution for live acquisition, and methods for performing live analysis on systems without interrupting the execution sequence.

In the case of analysis of quiescent systems the process of data acquisition from storage media is generally well-accepted at this time, and has the useful characteristics that digital copies of storage media can be made while preserving the original media in an unaltered state, and that the data acquisition and analysis process can be repeated at any later point in time with (hopefully) the same results. It is far less clear that such assurances can be made for data acquisition on non-quiescent systems, as actions taken by the investigator may change the state of the target system, and the dynamic state observed by the examiner may not be reproducible, preventing repeated analysis of the same state. As a result, there are certainly legal questions that must be considered as part of the research effort into the analysis of non-quiescent systems.

4.1.5. Media Types. The field of Computer Forensics has evolved into Digital Forensics. This change of name is not merely cosmetic, but indicates the wide range of digital devices that are often part of an investigation. However, while devices such as phones, digital media players, and game consoles may harbor relevant information, there are some significant challenges associated with forensic analysis of such devices. Cell phones are perhaps the most diverse, as they tend to have no standard interface, either at the hardware or software levels, essentially making the analysis process unique to each device model. Furthermore, forensic tools often cannot handle new or less commonly encountered devices, leaving an investigator to either develop custom tools, or lose the opportunity to examine the device. In addition to the number of incompatible devices of a particular type, such as cell phones, the number of device types, especially integrated devices, is also growing rapidly, as shown in section 4.1.2.

4.1.6. Control Systems. Process control systems (SCADA Systems) generated much discussion as an area that the security community recognizes as a security threat, but not yet perceived by industry to be as much of a threat. As a result, this field lags behind most technical fields in the area of security. These systems are potentially more vulnerable to attack and more likely to need associated digital forensics capabilities. Unfortunately, most process control systems were not built to track their processes, but merely to control them. As a result, many significant research and development categories were identified under this area. The participants acknowledged that initial focus is a primary concern of this area.

Subcategories identified in this area include the collection of evidence in the absence of persistent memory, hardware-based capture devices for control systems network audit trails, honeypots for control systems as part of the investigatory process, radio frequency forensics (900MHz), and intrusion detection systems for control systems. In addition to research related to digital forensics, the participants discussed the necessity for a development agenda for process control systems that includes security during all phases.

A research and development agenda for this area is being undertaken by a subset of the workshop participants in order to provide input into how this problem can be addressed as a priority item for protecting our critical infrastructure. The results will likely be reincorporated into a single digital forensics hierarchy diagram that captures the research agenda into digital forensics.

4.2. Education

The education research agenda was difficult to approach as it is challenging to separate the *research in education* needs, where we are conducting research to help identify better ways to educate our constituencies with respect to digital forensics, from education and training needs. Research in education for digital forensics will help us to identify the educational methodologies, materials, and environments that will assist educators in meeting the educational and training needs of their diverse constituencies. The categorization of educational

research needed to advance the field of digital forensics is being undertaken by a subset of the workshop participants.

4.3. Legal Issues

The legal issues associated with digital forensics were also considered an overarching theme that would be difficult to incorporate into a single hierarchy. Identified legal categories include Constitutional Law, Property Law, Contract Law, Tort Law, Cybercrime, Criminal Procedure, Evidence Law, Cyber War, as well as special issues. Beyond the categories listed, there are additional complications associated when the arena is extended to an international playing field. Thus International Law, is a secondary overarching legal issue that merits further research. This work is also being undertaken by a subset of the participants.

5. Future Work

Enumerating research categories and problems was a challenge. The team made significant progress in a short time. The organization of categories into a formal hierarchy proved to be more problematic. The initial categorical organization shown in Figure 1 is a starting point, but is by no means complete, nor does it represent an optimal organization of the categories presented.

The separation of Legal Issues and Education from the initial classification system provides the smaller working groups with a more approachable problem to solve as well as the potential to organize the research areas in a manner consistent with the area being investigated. In addition, the separation of Process Control Systems, at least in the preliminary phase, provides for more flexibility in expanding this category of research (and development), and allows a focused team the opportunity to begin to address the many security issues facing this category that is such a vital part of our critical infrastructure.

Ultimately, like all other fields of research, the field of forensics springs from basic principles. As an example, the results of forensic analysis are analogous to telling a story. The manner in which the story is told depends on its audience. An audience of technical experts will not require the same depth of explanation because much of the details of the story will be obvious to them; but a non-technical audience will require more exposition throughout in order to comprehend the material. An audience drawn from the legal community will require detailed information on how the evidence used to support the story was gathered, and how it was protected (i.e., the chain of custody), whereas a technical audience might not consider that as important as what the evidence reveals.

Three lines of inquiry emerge from the basic principle of forensics telling a story. They are:

1. How has the data been collected?
2. How has it been interpreted?
3. How has the resulting interpretation been conveyed to its audience?

5.1. Data Collection

Analysts gather data in a variety of ways. They can monitor a network or set of hosts or devices. They do so at various locations in the network (at the gateway or firewall, for example), or on the system or device in question. The data itself can be obtained in real time (as is typical with intrusion detection systems), as part of a post mortem analysis of a system (as is typical from log files of a crashed or compromised system), or the analysts can anticipate the type of data they will need, and instrument the system to record it. In the last case, the absence of data may be as reveling as its presence.

5.2. Data Interpretation

As the data is gathered, the analysts must interpret it. A variety of technical and non-technical factors come into play. As an example, the processing rate for monitoring networks is critical to detecting problems. The classic example is monitoring gigabit networks, where current technology cannot record all traffic without affecting the network's transmission speed. Similarly, a lack of understanding of how systems work can lead to errors. A stealth attack, in which attackers add packets that they know will pass monitors but never reach their final destination, will remain undetected unless the analysts understand that the record of traffic at the monitoring point must be interpreted in light of what portion of that traffic will reach the end point [2].

Other technical questions arise from the context of the work. For example, on most Linux systems, executing the program “mail” indicates that the user is reading mail or sending a letter. But this assumes that the search path for the user has the system directory containing “mail” before any other directories containing a program called “mail”. It also assumes that the user’s shell’s notion of white space is the traditional one, and that the environment is compatible with that of the analyst.

Worse are the non-technical factors involved in the analysis of the data. For example, a particular password may look very difficult to guess if the analyst speaks only English, but a native Russian speaker might immediately recognize it. In electronic mail or messages, a non-native speaker may misuse English words, leading to misinterpretations or misunderstandings. One who is not acclimatized to the culture of multilevel security may repeatedly attempt to violate the rules through ignorance rather than malevolence. These factors are critical to providing an accurate story.

5.3. Conveying the Interpretation

The third line of inquiry is how to convey the results to the intended audience. A lay jury may be impressed by a statement like "the use of the stealthy injection of packets clearly indicates the sender was trying to evade the detection mechanisms" but in fact, technologically, the statement embodies a number of assumptions that must be validated before such a claim can be made[1]. The technical sophistication of the audience is critical here.

This suggests asking what the goal of some forensic analysis is, and performing a stepwise refinement of the goal. As the goal is refined, the principles begin to emerge. We can then apply these principles to a wide variety of situations, especially those no-one anticipated. This ability distinguishes academic education from training, and is crucial to the advancement of science—the questioning of assumptions, and the development of scientifically rigorous and repeatable experiments that validate the results. Failure to do so raises questions about the integrity and accuracy of the results. This principles-based approach to refining the research agenda may prove to be a foundational categorization onto which the specific research sub-categories can be overlaid.

6. Conclusions

The categories and topics described in this paper clearly demonstrate the need for top-down research in digital forensics. Recent advances in the field provide both challenges and opportunities. The key to overcoming the challenges, as well capitalizing on the opportunities, will be timely research. The research areas outlined in this paper are identified as starting points and both the research agenda and the associated technologies will evolve as progress is made. Four teams are focusing on refining subsets of this problem (Process Control Systems, Legal Issues, Education, and Research) including mapping the categories using a principles-based approach. The authors suggest that this work could serve as the foundation for further advances in this area.

[1] Some specific points are whether the injection of packets occurred between the sender and the monitoring point—in which case someone other than the sender was doing the injecting; how the analysts determined that the packets were being "stealthily injected" rather than merely "injected" and what that means; and whether the sender's software inserted the packets without her knowledge, as might happen if the sender's system was compromised, for example by a Trojan horse.

7. Acknowledgements

The authors would like to thank R. Vaughn for inspiring this work, the Digital Forensics Working Group participants for their contributions and time, as well as the many fine researchers and practitioners who have contributed to the evolution of digital forensics.

8. References

[1] Pollitt, M., Nance, K., Hay, B., Dodge, R., Craiger, P., Burke, P., Marberry, C., and Brubaker, B. "Virtualization and Digital Forensics: A Research and Education Agenda," J. Digital Forensic Practice, vol. 2, no. 2, 2008, pp. 62–73.

[2] T. Ptacek and T. Newsham, "Insertion, Evasion, and Denial of Service: Eluding Network Intrusion Detection," Technical Report, Secure Networks, Inc.. Calgary, Alberta, Canada (1998).