Chapter 1

SOME CHALLENGES IN DIGITAL FORENSICS

Eugene Spafford

Abstract This essay discusses some of the principal challenges facing the emerging discipline of digital forensics. Most of the challenges have a scientific basis—understanding the needs and limitations caused by changes in the scope and pace of information technology. Others are engineering in nature, requiring the construction of new software and hardware to enable the collection, retention and examination of potential digital evidence. All of the challenges have administrative and legal frameworks within which they must be addressed, and the limits and structures imposed by these frameworks must evolve and be shaped by science, engineering and practice.

Keywords: Digital forensics, research challenges, science, engineering, practice

The use of information technology continues to grow at a rapid pace. Sometime in 2005, the world population using the Internet exceeded one billion [6]; estimates are that it will double in less than a decade. Computing devices are being used to communicate, bank, shop, operate businesses, interact with governments, learn and seek entertainment. Simultaneously—and not unexpectedly—criminal activity has risen along with the increase in the user population. One estimate puts global losses from cyber crime at more than $105 billion per year [2]. Waste and abuse may well match or exceed these figures.

Misuse of information technology resources is a major problem that cannot be addressed solely by better security technologies. The situation is complicated by the need for backwards compatibility, lack of user awareness and education, limits to known technologies, and the massive base of installed infrastructure with little or no support for security. The prospects are dim for near-term solutions to many existing problems [7].

If we cannot reengineer our information infrastructure to be completely protected, then we need to address the problems of cyber crime and abuse after they occur: by investigation and corrective action, including application of remedial measures, as well as legal and administrative sanctions. This requires comprehensive tools and technologies for investigation that can be trusted to provide accurate, precise results. It also requires competent investigators who are trained in these tools and technologies so as to draw the correct conclusions.

Digital forensics is a relatively new field. Until the mid 1990s, the only public instances of code and log analysis involved detecting intrusions and misuse, or perhaps making some incidental observations about a potential online miscreant, such as a malware author [5, 8]. In 1992, Cornell juniors David Blumenthal and Mark Pilgrim were arrested for writing and releasing the MBDF virus that targeted Macintosh computers [3]. They were identified by a group of anti-virus researchers (including me) after examining the virus source code and system logs. That same year, the first formal paper about software forensics was written and presented [9]. Thus, we might identify 1992 as the year when digital forensics began to emerge as an identifiable field. However, it is still in a nascent phase where we are trying to identify the scope of the field and to decide what to call it: computer forensics, cyber forensics, digital forensics, digital investigations and so on.

Whatever we call this field, it is, nevertheless, useful to examine its scope and current status. One promising classification is to consider the continuum of science, engineering and practice. Each has its own set of unique challenges and needs. There are no clear demarcations between the three areas, but all three are important and necessary components that need to interconnect and communicate.

- **Science:** We can think of science as the formal investigation and documentation of principles, limitations and structure of a field. Science is performed by formulating hypotheses that can be confirmed or refuted, and then conducting carefully designed experiments or analyses. The outcomes must be meticulously documented and then presented to the community, so others can recreate the results and build on them.

  The science of digital forensics is still quite limited. We have seen only a few formal models of how digital forensic investigations are structured and conducted (see, e.g., [4]). We still need to understand the limitations and capabilities of forensic investigations. We also need to examine how the potential for digital forensics can be expanded by designing specialized forensic support within new systems (see, e.g., [1]).

- **Engineering:** Engineering can be viewed as the development and application of tools and procedures to solve real problems with known parameters. For example, engineering addresses the questions of how to reliably find data on known (and new) media, and how to reliably distinguish between multiple alternatives in the recreation of incidents. This has been an area where there has been significant development, although not all of it has been formally conducted and published, especially efforts that have been undertaken by commercial entities.

- **Practice:** The practice of digital forensics does not involve the creation of tools or research into technical issues. Instead, it involves training and the application of known tools and techniques within established limits to address real needs. Practice is a major component of digital forensics and it requires better methods for training and setting standards.

The development of digital forensics has mirrored that of many other disciplines. Problems that arise are initially addressed using tools and techniques developed to solve different but similar problems. Some of the first tools applied to digital forensics were developed for system administration and software debugging. As needs grew, engineering expertise was brought to bear to develop new tools and techniques, leading to specialized procedures and training. The synergy between engineering and practice continues, but new challenges and the anticipation of future needs are now drawing more scientific efforts.

In my keynote lecture at the 2001 Digital Forensics Research Workshop I outlined some of the research challenges I saw in the field of digital forensics. Most of these challenges remain and some have expanded; I do not believe that any of them have been adequately addressed as yet.

New challenges are also presenting themselves as technologies advance. Many of these are related to the incredible growth of storage capacity and the pace of the growth. In 1995, there were about 200 terabytes of storage connected to the Internet; this amount of storage is now accommodated in about 100 commodity computers. However, as reported by IDC and recounted in a 2004 posting to Dave Farber's IP list, worldwide storage connected to the Internet exceeds 30 exabytes—a 150,000-fold increase in capacity in only one decade. This growth in storage is continuing, which means that more information must be protected and more information must be analyzed to discover evidence about incidents.

In the following, I list several open challenges that need to be addressed. Some may be solved by engineering advancements; others require significant scientific investigations.

- How do we copy terabyte (and larger) storage media rather than confiscating them? There is a requirement to obtain and hold evidence, but it is inappropriate to inconvenience innocent victims and third parties who may require the media for business operations. However, the capacity of these media and the available transfer rates to other storage devices make bit-for-bit duplication a major challenge.

  Perhaps we must redefine the term "everything needed" when evidence is collected to reduce the amount of data that needs to be copied. Or, we might combine *in situ* examination with copying to reduce the scope of what needs to be copied.

- How do we image large, active disk farms dynamically? Victims and third-party providers may have evidentiary material on their media, but shutting their systems down to make a copy can greatly inconvenience them. Imagine asking Amazon.com or eBay to discontinue service while their drives are being copied!

  In addition to the mechanics of copying, we also need to understand how the copies relate to the media at the time of the crime, and we need to present that information according to acceptable standards of evidence.

- Where can storage reside that may contain evidence? In the last few years we have seen an increase in the use of USB "thumbdrives," cell phones, digital cameras, PDAs, remote storage devices and removable media. Understanding the scope and range of storage continues to be a challenge in forensic investigations. The emerging use of *ad hoc* networks, RAID over network servers, and long-term storage in appliances and home media will blur the notion of "local storage."

  Furthermore, the use of open 802.11 networks in neighborhoods, cable modems and unsecured, unpatched systems means that perpetrators can store data of interest on a number of systems that are not obviously under their control. Not only do we need to be able to examine the evidence on these systems, but we must exclude quickly the actual owners of the systems as suspects.

- How can we trust audit trails? There is always the possibility that an intruder (or his software) may edit or delete the audit trail on a computer, especially a weakly-protected PC. Furthermore, over the last few years we have seen increasingly sophisticated rootkits that dynamically modify the kernels of running systems to hide what is happening—or even to produce false results.

- How do we accurately reconcile evidence collected from multiple machines that may not have accurate clocks to appropriately sequence events? This becomes even more important as more machines are involved, including machines in different time zones.
- How do we deal with non-determinism? A number of systems use randomness or asynchronous events in their processing. If we need to understand how something happened, how do we recreate the relevant events and their timings? How do we accomplish this if the input comes from *ad hoc* networks of sensors that cannot be "reset" to an earlier state?
- How do we cope with the changing nature of what needs to be investigated? Instead of simple data and image files, we are now seeing video, audio, GIS material, VoIP systems, sensor net data, SCADA systems and more. What are the standards for terminology, collection and representation that will allow investigations to be conducted accurately on these systems?
- What are the limits over time of what we can do? If we perform forensic examinations of backups and mirror sites, how much can we accurately conclude?

Several of the challenges have a scientific basis—understanding the needs and limitations caused by changes in the scope and pace of information technology. Others are engineering in nature, requiring the construction of new software and hardware to enable the collection, retention and examination of potential digital evidence. All of these challenges have administrative and legal frameworks within which they need to be addressed, and the limits and structures imposed by these frameworks need to evolve and be shaped by the science of what is possible, by the availability of engineered solutions and by disciplined practice.

In Asimov's 1956 short story, *The Dead Past*, a scientist helps develop an inexpensive "time viewer" that allows one to see incidents from the past. After much thought, his superiors decide that no one should know about their discovery. The scientist was stunned. He believed that the world would benefit from the technology: old crimes could be viewed and solved, the causes of accidents could be traced, historical disputes could be settled. The scientist acknowledged that for certain incidents, such as those involving religious figures, disclosures from the past might be traumatic, but they would be therapeutic in the long term.

The scientist made copies of the blueprints and sent them to newspapers and leaders around the world. Only after the blueprints went out did his superiors find out what he had done. It was then that they

pointed out that the past is not simply long dead events, but also very recent incidents. By tuning the viewer to half a second in the past, one could snoop on what anyone was doing *right now*. The tragedy was that those who wanted to snoop on others could do so. The ability to see the past resulted in the loss of privacy for all.

This is an apt parable for the digital forensics community. Our ability to analyze data from the past makes it possible to examine the current behavior and activities of those we might wish to monitor. This capability presents opportunities, but also new responsibilities.

Clearly, there are significant ethical issues that must be addressed concomitantly with technological advancements in digital forensics. These issues need to be identified, resolved and articulated. As digital forensic technology improves, it may well become easier to discover details about people's lives, loves and activities, especially as more information is stored online and maintained indefinitely. Digital forensic practitioners have a duty to identify the guilty and exonerate the innocent in issues of misbehavior and crime, but they should also ensure that they preserve the privacy of all parties, principals as well as incidental contacts. Privacy, once violated, is difficult—if not impossible—to restore.

Digital forensics is an engaging, vibrant field. Many challenges exist, but opportunities abound to innovate and make a difference. We ought to refrain from needless arguments about what to call the discipline and look beyond the next set of problems to address. We should consider how we want to be known and what roles we should play. Technical challenges may shape what we do, but come what may, there will always be a role for sound human judgment.

References

[1] F. Buchholz, Pervasive Binding of Labels to System Processes, Ph.D. Dissertation, CERIAS Technical Report 2005-54, Department of Computer Sciences, Purdue University, West Lafayette, Indiana, 2005.

[2] Cable News Network (CNN), Record bad year for tech security (money.cnn.com/2005/12/29/technology/computer_security/index.htm), December 29, 2005.

[3] J. Carmona, Computer virus traced to Cornell students, *The Cornell Daily Sun*, February 25, 1992.

[4] B. Carrier, A Hypothesis-Based Approach to Digital Forensic Investigations, Ph.D. Dissertation, CERIAS Technical Report 2006-06, Department of Computer Sciences, Purdue University, West Lafayette, Indiana, 2006.

[5] T. Longstaff and E. Schultz, Beyond preliminary analysis of the WANK and OILZ worms: A case study of malicious code, *Computers and Security*, vol. 12(1), pp. 61-77, 1993.

[6] Miniworld Marketing Group, Internet world stats (www.internet worldstats.com/stats.htm), June 1, 2006.

[7] President's Information Technology Advisory Committee (PITAC), *Cyber Security: A Crisis of Prioritization*, National Coordination Office for Information Technology Research and Development, Arlington, Virginia, 2005.

[8] E. Spafford, The Internet worm program: An analysis, *Computer Communication Review*, vol. 19(1), pp. 17-57, 1989.

[9] E. Spafford and S. Weeber, Software forensics: Can we track code to its authors? *Computers and Security*, vol. 12(6), pp. 585-595, 1993.