Practitioners’ Challenges Panel:  
The Challenges and Opportunities of Forensic Investigation over Large Data Sets

Timeliness vs Precision vs Comprehensiveness

Nan Zhang, The George Washington University, nzhang10@gwu.edu  
Yong Guan, Iowa State University, yguan@iastate.edu  
Michael M Losavio, University of Louisville, michael.losavio@louisville.edu  
Peter Vasquez, VTX Communications, pvasquez@vtxc.net  
Robert M. Nissen, National Computer Security Center, r.nissen@radium.ncsc.mil  
Edward Talbot, Sandia National Laboratory, edtalbo@sandia.gov  
Vassil Roussev, University of New Orleans, vassil@cs.uno.edu

I. Executive Summary

The practitioners’ challenges panel addressed present and future challenges to the effective use of digital forensics. The goal is to understand the needs and real challenging/hard problems the practitioners have and are facing, and achieve consensus on the research opportunities and directions the DF community should focus on.

A starting point is discussion of how the digital forensics research community can respond to a pressing need from practitioners on conducting forensics over very large data sets (e.g., a storage device containing terabytes of files, or high-volume high-speed network transmitting gigabytes of traffic per second) in an efficient and on-demand manner (resp., reaching pre-defined error bounds within minutes). Timely and error-bounded, “first-response” forensics is urgent needed in many critical applications, such as border patrol, enterprise surveillance, and civil/criminal investigations.

One motivating challenge here is the growing discrepancy between the advancement of transmission and storage capacity and performance: In the past few decades, storage capacity increases about 60% per year, while storage performance such as throughput and latency see much slower improvement. The traditional two-step process, i.e., whole-device imaging with comprehensive analysis of the image cannot cope as the amount of time required is too long to be practical for border patrol and enterprise surveillance, which require an error-bounded forensic analysis of very large data sets (e.g., terabytes of data) within matter of minutes.

The objective of this panel is three-fold:
1. To identify a list of time and error bounds, in terms of completeness, false alarm rate, amount of human intervention, etc., for real-world first-response forensics applications, from a practitioner’s perspective from recovery response to prosecution.

2. To identify a set of promising technical leads that the research community may follow to address these time and error bounds.

3. To identify the impact of first-response forensics on decision-making, policy issues, etc., from a practitioner’s perspective, from recovery response to prosecution. The inevitable change of forensics techniques to deal with very large data sets brings along new challenges on the policy front. For example, given the absence of full-device imaging, how should one decide whether the data retrieved so far demonstrate sufficient evidence warranting further investigation? Also, how to address the legal issues raised by possible false alarms?

II. Questions for Discussion:

What are the future needs of law enforcement given the increasing sophistication and growth in cyber crime, particularly involving ubiquitous digital devices and transnational connectivity?

What are the concerns of private industry, particularly telecommunications and device manufacturers, with implementing forensically sound collection of evidence?

How should whole-device imaging and first-response forensics complement each other in a forensics investigation? Just like in medical emergencies where first-responders must hand over control to hospital care at some point of time, when should a forensics investigator determine that enough evidence (“probable cause”) have been collected by first-response forensics so it is time to start a full-blown whole-device imaging process?

What are the real-world forensics applications which have the most pressing need for first-response forensics?

How “fast” is fast? What is the maximum amount of time tolerable for real-world applications that require first-response forensics?

How “accurate” is accurate? What is the maximum amount of false alarms tolerable for real-world applications? In terms of false negatives (completeness?), should first-response forensics be given more leeway (or how much) than whole-device imaging, just like the Good Samaritan Laws which protect medical first responders?

What are the consequences (pros and cons) of applying methods like sampling to forensics? Would the paradigm shift from deterministic to probabilistic answers be accepted by forensics practitioners? For example, given that sampling has seen significant success in other areas, how to push for multi-disciplinary efforts (from forensics, storage, database, data streaming communities, or even from statisticians) to investigate the problem of first-response forensics?
III. Summary of Panel Discussions

The panel started with organizers outlining the vision of the panel and the potential questions for discussions, followed by the panelists stating their most critical concerns regarding forensics investigations in practice. A number of issues were raised, including

- the impact of forensics investigations on the business world – e.g., how can private industries cope with policy issues related to forensics investigations, how can device manufacturers implement devices that can produce forensically sound collection of evidence, etc.;
- the lack of clear specification on the requirements of forensics investigations – in particular, quantitative performance measures that academic researchers and industry developers could use to evaluate the utility of their results/products;
- the challenges brought by the advancement of technology to real-world forensics investigations – a particular issue being raised multiple times here is scalability, e.g., how to deal with a storage device containing terabytes of data, or high-volume high-speed network transmitting gigabytes of traffic per second.

After the opening statements a number of topics were extensively discussed (and sometimes debated over). One such topic is the gap between prototypical solutions provided by academic research and the quality/robustness required by real-world forensics practitioners. A number of critical concerns raised in the opening statements, e.g., the scalability issues, how to judge forensic soundness of data collection, etc., are related to this topic. Potential solutions suggested by the panelists include calling upon the community to collaborate with practitioners on solving real-world problems, calling upon practitioners to increase interactions with academic researchers, increasing research support from federal government on forensics research, etc.

More technically, a topic that was extensively discussed is how to address the scalability issue, especially in a real-time fashion. Again, multiple panelists raised the issue of lacking proper performance measures and standards - which should be best set by real-world forensics practitioners rather than academic researchers providing technical solutions. To address the issue, some panelists suggested learning from the security community on their tools featuring real-world processing and protection, with the vision being that the development of real-world forensics tools can also return the favor by improving the performance of security protection.

Finally, the panelists discussed the challenges of integrating forensics research into training and education. A key challenge discussed here is cost issues related to teaching forensics in a vendor-neutral manner. The panelists agreed that negotiating with forensics software vendors or collaborating with industrial companies may provide a potential solution to the problem.