Chapter 1

CALIBRATION TESTING OF NETWORK TAP DEVICES

Barbara Endicott-Popovsky, Brian Chee and Deborah Frincke

Abstract Understanding the behavior of network forensic devices is important to support prosecutions of malicious conduct on computer networks as well as legal remedies for false accusations of network management negligence. Individuals who seek to establish the credibility of network forensic data must speak competently about how the data was gathered and the potential for data loss. Unfortunately, manufacturers rarely provide information about the performance of low-layer network devices at a level that will survive legal challenges. This paper proposes a first step toward an independent calibration standard by establishing a validation testing methodology for evaluating forensic taps against manufacturer specifications. The methodology and the theoretical analysis that led to its development are offered as a conceptual framework for developing a standard and to "operationalize" network forensic readiness. This paper also provides details of an exemplar test, testing environment, procedures and results.

Keywords: Network forensics, aggregating tap, calibration, baseline testing

1. Introduction

This paper presents an approach – derived from courtroom admissibility standards – for calibrating low-layer network devices employed in collecting data for use in courtroom proceedings. The collected data may be used to prosecute malicious conduct on networks or to seek legal remedy for (or defend against) accusations of network management negligence. While we specifically discuss our approach in the context of aggregator taps, it can be generalized to more complex devices. The model is offered as a first step towards filling a void created by manufacturers who provide general specifications for taps and switches that

Please use the following format when citing this chapter:

Endicott-Popovsky, B., Chee, B., Frincke, D., 2007, in IFIP International Federation for Information Processing, Volume 242, Advances in Digital Forensics III; eds. P. Craiger and S Shenoi; (Boston: Springer), pp. 3-19.

collect forensic data, but offer few guarantees about the actual behavior of these devices.

Several factors are responsible for the lack of calibration regimes for network forensic devices. In an intensely competitive market, vendors often consider the architectural details and precise behavior of their data-gathering network devices as proprietary. Purchasers select and employ these network devices primarily for troubleshooting as opposed to gathering evidence that would withstand courtroom scrutiny [18]. Furthermore, standards and precedents needed to establish the validity of data-gathering devices in legal proceedings are only beginning to emerge. Consequently, even vendors who are interested in meeting the requirements for forensic soundness do not have a set of best practices for device testing and validation.

This paper deals with calibration testing of network devices. In particular, it focuses on the role of calibration in establishing a foundation for expert testimony. While much consideration has been given to recovering forensic data and using it as digital evidence, little attention has been paid to calibrating the hardware devices used to capture network traffic and documenting how they behave "in the field." With information technology and cyber-enabled activities becoming ever more important factors in legal proceedings [17], the consequences of failing to evaluate and validate the behavior of network forensic devices could lead to inadmissible evidence and failed legal action.

2. Calibration Testing

Lord Kelvin, a giant in the field of measurement science, eloquently described the role of calibration in an 1883 lecture to the Institution of Civil Engineers [3]:

> "I often say that when you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot express it in numbers your knowledge is a meager and unsatisfactory kind; it may be the beginning of knowledge but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be."

Calibration is "the comparison of instrument performance to a standard of known accuracy in order to determine deviation from nominal and/or make adjustments to minimize error" [3]. Calibration is conducted when there is a need for confidence that a piece of equipment performs as intended. Calibration testing can be conducted by professionals performing formal tests at a standards laboratory, by internal experts using an in-house metrology laboratory, or even by users per-

forming verification testing to ensure that the procured equipment meets manufacturing specifications.

A formal calibration test is typically accompanied by documentation that provides traceability to the standard used, the periodicity of the calibration (the interval between tests that provides a continuity of confidence that an instrument is performing within an acceptable band of tolerance), and an expression of imprecision that takes into account the potential for error in the test and/or test equipment [2]. A simple example of calibration is the testing of weights and measures to ensure that a customer purchasing a pound of flour from Merchant A is, in fact, purchasing a full pound.

In the forensic sciences, calibration is employed to ensure that instruments used in the collection and analysis of evidence can be relied upon [13]. For example, radar guns used by police to clock speeders must be calibrated according to state or local laws. These laws also establish the calibration regimes that specify the periodicity of testing, testing documentation requirements, and the duration that records must be maintained [22].

Evidence of radar gun calibration provides the foundation for testimony that an individual was speeding [13]. Absent such evidence, speeding charges may be dismissed [22]. Also, if the testing regime is not followed precisely, a valid defense might be: "*How do you, Mr. State Trooper, know the radar gun was working properly?*"

By analogy, if devices on a network are used to collect electronic evidence, then the performance of the devices must be understood and documented if the collected evidence is to survive similar courtroom challenges. For example, if, at the time of a network incident, the behavior of the forensic device cannot be characterized as adequate for recording relevant network data, a defense could be mounted that it was "someone else" who violated the system. Hypothetically, it could be alleged that significant numbers of packets could have been dropped. A defense attorney might ask: "*Why conclude that my client is responsible when exculpatory evidence may be missing?*"

3. Frye and Daubert Standards

Certain landmark court cases have established rules for admissibility, which ensure that scientific forensic testimony is "relevant, material and competent" [13]. *Frye v. United States* [23] established the standards under which judges should accept expert testimony. *Daubert v. Merrell Dow Pharmaceuticals, Inc.* [24] established that Rule 702 of the Federal

Table 1. Frye vs. Daubert standards.

Frye Standards	Daubert/Kumho Factors
Is the approach sufficiently established?	Has the technique used to collect evidence been tested? (Or can it be tested?)
Has the technique gained general acceptance in its field?	Has the theory underlying the procedure or the technique itself been subjected to peer review and publication?
Does it require study/experience to gain special knowledge?	Does the scientific technique have a known or potential rate of error?
Does expertise lie in common experience and knowledge?	Do standards exist, along with maintenance standards, for controlling the technique's operation?

Rules of Evidence supercedes Frye. This decision was further enunciated in *Kumho Tire Co. v. Carmichael* [25].

Table 1 summarizes the differences between Frye and Daubert. The main difference is that Frye establishes the general acceptance standard and some rules for the admissibility of evidence, while Daubert relaxes those rules and allows judicial discretion. Frye has tended to support the exclusion of evidence rather than its inclusion, especially when the issue is in doubt. On the other hand, Daubert allows judges more leeway in accepting evidence obtained by the application of new technologies.

Expert testimony is generally used in one of two ways: (i) the expert explains evidence in a way that a jury can understand, or (ii) the issues are so complex that only the expert can understand them; therefore, expert believability is based on trust established with the jury [13]. In the second instance, the opposing counsel may seek to discredit the expert witness and his/her testimony. One strategy is to challenge the testimony's foundation by asking questions such as: "*How do you know this?*" "*How can you say this?*" "*How can we believe the validity of what you say?*"

Lack of personal knowledge of the science behind the testimony does not bar a jury's trust. However, the greater the trust expected of the jury, the greater the burden to provide competent foundation that sup-

ports both the credibility of the witness and the equipment used to gather evidence [13]. This foundation does not identify the guilty party; instead, it provides the basis for believing the expert's testimony: "*Here is the tool I used.*" "*Here is the data that it describes.*" "*This is why the tool works.*" Calibration is part of that description and speaks to the reliability and predictability of the tool.

In a review of several hundred pages of digital forensics testimony involved in cases in the Pacific Northwest from 2000 to present, we discovered that the technical competence of the evidence and the questioning of expert witnesses, ranged from minimally competent to highly professional. In the cases examined, experts represented both the prosecution and the defense. Note that while federal government experts are required to have demonstrated levels of expertise (usually manifested by certifications), the expertise possessed by local law enforcement and defense experts ranged considerably from case to case. In one instance, an uninformed defense "expert" testified there were "100 bits in a byte" and calculated network traffic flow based on this erroneous value [11]! In some cases, a modest, even deficient, understanding of technology was sufficient to introduce "reasonable doubt" in felony trials and persuade juries to acquit the defendants [11].

The Frye and Daubert tests provide some protection against the use of bogus scientific evidence and expert opinion. But in the end, particularly under the Daubert standard, the task of challenging inexact science falls on the attorneys in the courtroom. While the legal bar's understanding of digital forensics is usually very limited, often allowing technically incompetent evidence to go unchallenged, the state of the practice is improving [13]. This makes the establishment of a proper foundation for network evidence gathering even more important. In fact, how other technical devices for gathering evidence (e.g., radar guns, mass spectrometers and gas chromatographs) are maintained and presented in courtroom proceedings provide insight into what will be expected of devices used to collect network evidence.

One option is to suggest that only properly calibrated collection methods be employed, but this is not practical. Organizations are already using switches with span ports and aggregator taps with monitoring ports to capture network forensic data [15, 16]. These devices are rarely, if ever, calibrated to the extent that a proper foundation can be laid for their data-gathering accuracy. To complicate matters, marketing literature for these devices claims "forensic capability" without defining what is meant by "forensic" [5]. Network incident response personnel frequently use the term "forensic" inaccurately to mean troubleshooting (i.e., determining what happened so that the system can be fixed or re-

stored), as opposed to collecting data that meets courtroom admissibility standards.

Sommers [18] calls this forensics with a "little f" in contrast with Forensics with a "capital F," which seeks, in addition, to determine who is responsible. While it is true that span port and aggregator features were designed to support troubleshooting (forensics with a small "f"), there is evidence that some entities are using these devices to collect data for legal proceedings (forensics with a capital "F") [4, 16].

Given the likelihood of increasingly sophisticated examination of expert testimony, courtroom admissibility requirements are expected to become an important consideration for network devices (although at present they do not appear to be quite so important). This provides a window of opportunity for developing standards, tests and regimes for network evidence gathering devices.

4. Baseline Testing

Absent vendor certification, it is necessary to develop a suite of standard tests for validating manufacturers' specifications for devices used for collecting forensic data. The testing would allow network traffic collected for evidentiary purposes to be described competently during expert witness testimony. A suitable approach is to adapt network baseline testing techniques for this task.

Baselining is defined as "systematically viewing network point-to-point data flow to analyze communication sequencing, extract accurate metrics, develop a technical synopsis of the network and make recommendations" [12]. Quantitative statistical measurements are employed to identify key network issues relevant to supporting the mission of an enterprise [12]. Baseline testing can be used to provide assurances that a network is stable and operating reliably and also to support decision-making, e.g., the need for investment in increased network capacity. The baselined period is usually one week, but some enterprises might experience monthly, quarterly and/or annual peak network loads, requiring the analysis of additional baseline performance data [12].

Baseline testing typically employs devices such as protocol analyzers to reveal network throughput performance. At a minimum, available capacity and utilization are determined; the latter is defined as the amount of capacity used by a certain network segment over a specific time interval that encompasses a typical business cycle [12]. Both average and peak utilization are of interest in determining the appropriate capacity.

While the goal of baselining is to provide a basis for managing network reliability, we believe that the testing approach can confirm the

capacity of devices used to collect network forensic data for evidentiary purposes. The test results coupled with an understanding of the baseline performance of the network segment where the device is operating would allow the forensic data gathered by the device to be characterized (e.g., providing a formal statement about data completeness at the time it was collected). For example, if a network typically runs at 20% capacity and if it is known from testing that the device functions at line rate capacity under this condition, then an expert witness might assert that the data collected from the device at that time is complete for all intents and purposes.

To provide additional detail, it is possible to use a suite of typical attacks to determine the likelihood that critical evidence is lost, including under peak loads. For example, in scenarios where there is considerable repetition of forensically-important packets (as in a DDoS attack), the probability that evidence is lost may be quite small. In other cases, such as a subtle attack involving privilege escalation, the probability of evidence loss may be higher. Under specific scenarios, it is important also to consider the perspectives of the prosecution/defense: Is there an advantage for the data gatherer not to gather all of the relevant data? Could a participant have interfered with the gathering of specific evidence to the extent of causing packet loss? Regardless of the scenario, a generalized model underlies the development of each calibration regime.

5. Calibration Test Model

We developed an exemplar case to create a model for calibration tests that could provide an adequate foundation for expert testimony. We identified a preliminary three-step process for devising a calibration test regime to address foundation challenges (Table 2). Subsequently, we applied the process to a case involving the use of a Net Optics 10/100BaseT Dual Port Aggregator Tap to gather forensic data.

5.1 Identifying Potential Challenge Areas

Given that organizations are already using switches and taps to capture forensic data, we limited the consideration of test subjects to switches with span ports and aggregator taps with monitoring ports. In theory, important data could be lost if these devices are oversubscribed. Furthermore, from the perspective of an expert witness, any lack of knowledge about the behavior of these devices could potentially damage his/her credibility with a jury.

Although more switches than taps are used for forensic purposes [16], we examined only taps in our study. Taps function as pass-through de-

Table 2. Validating foundation through network device calibration.

- **Step 1**: Identify a potential challenge area and perspective.
  - **Identify Challenge Area**: Ask how foundation testimony could be subject to challenge if data is lost.
  - **Perspective**: Expert testifying to completeness of data collection.

  Example Case: An oversubscribed forensic tap could drop packets. Inadequate characterization of the circumstances when this might occur could be used to challenge expert credibility, especially if comprehensive data collection is claimed.
- **Step 2**: Design calibration testing goals to support the challenge area (given the perspective).
  - **Goal**: Verify manufacturer's specification; describe device behavior.
  - **Perspective**: Expert witness determining whether it is reasonable to expect that all relevant data was gathered.

  Example Case: Validate tap capacity; determine when packets begin to be dropped.
- **Step 3**: Devise a test protocol.
  - **Purpose**: Ensure sufficient documentation/assurance that the test and test environment are appropriate for supporting expert testimony.
  - **Process**: Develop a "comprehensive" suite of stress tests that examine the behavior of the device in isolation.

  Example Case: An external laboratory was selected for testing the tap and a suite of tests created for a range of network traffic flows.

vices primarily at Layer 1 of the OSI model; switches function at Layer 2 or Layer 3 depending on built-in functionality. Typically, taps neither read/forward at the MAC address layer nor provide confirmation of a link state (both these conditions are necessary in a Layer 2 device). This makes taps simple to test and renders them neutral entities in the forensic data collection process, i.e., they merely pass the data stream without introducing latency. Because switches function at Layer 2 or Layer 3, they may introduce latency in packet forwarding, making the determination of data flow performance much more complex. In addition, manufacturers often treat embedded functionality as proprietary; this may require the switch architecture to be re-engineered to properly test its performance [4]. We selected the Net Optics 10/100BaseT Dual Port Aggregator Tap. This tap operates between Layers 1 and 2 because it incorporates embedded logic that aggregates duplex traffic and

forwards it to a monitor port. The tap was one of the first to be marketed as a "forensic" device, when 1 MB buffers were added to provide protection from network traffic spikes.

5.2 Designing Testing Goals

Our study focused on the ability of the Dual Port Aggregator Tap to keep up with data flow at the manufacturer-specified rate. Thus, the goal of the calibration test regime was to verify whether the tap could handle the combined traffic of a single full duplex link when traffic is at, or below, its 100 Mbps capacity [5] and to document any packet-dropping behavior. Any traffic exceeding the capacity was expected to fill the buffers up to 1 MB per side of the full duplex connection before any packets were dropped. We also sought to determine when packets would begin to drop. These goals became the basis for devising a test regime.

5.3 Devising a Test Regime or Protocol

Certain laboratory capabilities were required to ensure that calibration testing would help certify that the equipment and practices met acceptable standards. We began with two objectives for our test environment. The first was to obtain controllable traffic for pushing the tap limits. The second was to isolate tap behavior from other behaviors so that the foundation could be laid for tap performance in any environment, not just the test environment.

Isolating the tap proved challenging. Preliminary tests were unsatisfactory at separating the tap's behavior from NIC cards, the OS and the switch. The first test employed the `iperf` utility (v.1.7.0), part of the Knoppix-STD bootable Linux distribution. The second used the Fluke Optiview Online protocol analyzer. (See [8] for details about both tests.) The next option we considered was the Advanced Network Computing Laboratory (ANCL) facilities at the University of Hawaii, Manoa with multiple test beds set up specifically to eliminate external influences. We selected the Spirent Test Center, a high-speed, local area network test instrument designed to determine failure points in high speed networking equipment [19]. It is capable of generating and analyzing wire-rate traffic up to 10 Gbps, significantly higher than the vendor-specified 100 Mbps limit of the tap.

A series of tests to verify the most basic aspects of the manufacturer's specification was designed and adapted from baseline testing techniques. Since our challenge involved the loss of packets (or buffer overflow), these were the first (and only) tests we included as part of Step 3. The tests are

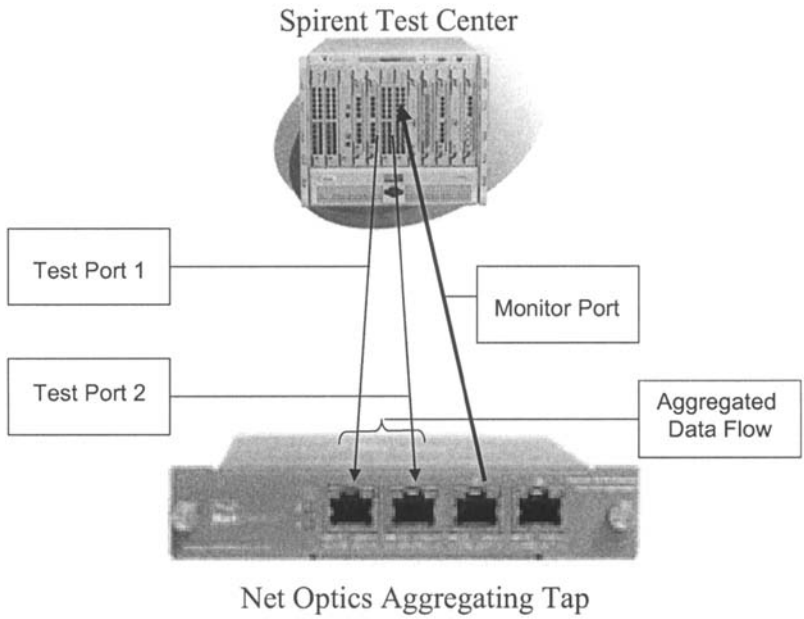

Figure 1. Aggregating tap test configuration.

by no means comprehensive, nor do they indicate anything unexpected about the device (this was not the intent). Rather, the tests demonstrate how a simple pre-courtroom calibration regime could be used to increase the credibility of witnesses who testify about data collected using the tap. In the future, we plan to consider malformed packets and to analyze tap behavior in the presence of common network attacks, with the goal of developing a comprehensive suite of exemplar tests.

The Spirent Test Center was configured as shown in Figure 1 to simultaneously transmit two equivalent data streams of homogeneously-sized UDP packets to Test Ports 1 and 2. The data streams were aggregated inside the tap and sent to the Monitor (forensic) Port where test data was collected. UDP packets were selected to eliminate latency introduced by the three-way TCP session handshake. Packet content was all 0's.

Four tests, each lasting 30 seconds, were conducted with data streams of consistently-sized packets of 64, 512, 1,500 and 9,000 bytes (Table 3). A fifth test involved random-sized packets (Table 3). For each packet size, tests were conducted across a range of data flow rates, expressed as a percent of the tap's 100 Mbps capacity. For example, the 50% and 51% data flow rate tests aggregated the rate of flow to the Monitor Port equal to 100% to 102% of capacity.

Table 3. Spirent Test Center test suite.

Packet Size (Bytes)	64	512	1,500	9,000	Random
Data Flow Rate (%)			**30 seconds**		
10%		X			
30%	X	X	X	X	
50%	X	X	X	X	X
51%	X	X	X	X	X
60%		X			
100%		X			
Packet Size (Bytes)	**64**	**512**	**1,500**	**9,000**	**Random**
Data Flow Rate (%)			**300 seconds**		
51%		X			
60%		X			

With a testing goal of verifying the tap capacity, we designed a calibration test regime by adapting benchmark test parameters from the Internet Engineering Task Force (IETF) Network Working Group RFC 2544 [6]. A series of duplex data streams from the Spirent Test Center were configured to send 512-byte packets (representative of average traffic) to the tap at varying rates of data flow expressed as a percentage of the tap's 100 Mbps capacity. Several tests were conducted for a duration of 30 seconds (marked "X" in Table 3). In addition, two tests were conducted for a duration of 300 seconds to observe any change in tap behavior. Oversubscription of the tap at a full duplex data flow rate of 51% coming from each side was confirmed with data streams of differently sized packets (64, 1,500 and 9,000 bytes, respectively) and randomly-sized packets [8]. Table 3 displays the series of tests that comprised the test regime.

Table 4 presents the results of the 512-byte packet test. Packets dropped when full duplex traffic was 51% or more of the tap's 100 Mbps capacity. At 10%, 30% and 50%, the Monitor Port received and forwarded the aggregated flow from Test Ports 1 and 2. At or above 51%, the Monitor Port was unable to forward the entire aggregated stream, verifying the 100 Mbps capacity specification at the Monitor Port as implemented by the manufacturer. The expected implementation of a 100 Mbps Monitor Port would forward the entire aggregated stream of 200 Mbps.

Figure 2(a) shows the graph of dropped packets for the 512-byte test series. Note the sharp rise at 51% of tap capacity indicating the data flow rate at which packets begin to drop. Figure 2(b) shows port average

Table 4. Spirent Test Center results (512-byte packets).

Traffic (% Tap Cap.)	Test Port 1 Transmitted	Test Port 2 Transmitted	Monitor Port Received	Dropped Packets
10	70,488	70,488	140,976	0
30	211,461	211,461	422,922	0
50	352,443	352,443	704,886	0
51	359,496	359,496	708,241	10,751
60	422,952	422,952	708,242	137,662
100	704,887	704,887	708,243	701,531

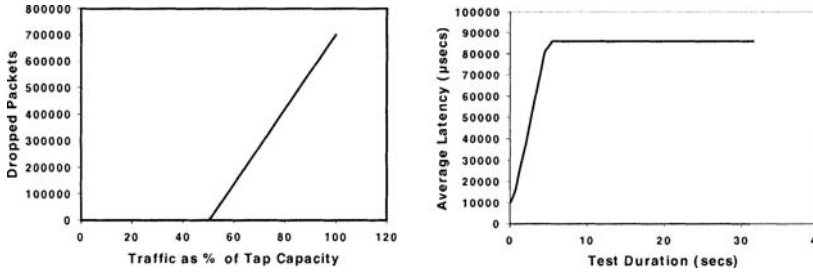

Figure 2. Test results: (a) Dropped packets; (b) Average latency.

latency through the tap. Examining the data, allowing for the one second test ramp up and one-half second period attributed to learning frames and clock variations in the console PC, the tap begins to drop packets four seconds into the test.

5.4 Predicting Packet Loss

Identifying where errors might arise in data collection is important for providing a foundation for evidence. Therefore, we have developed a probability curve for predicting packet loss in aggregating taps. The relationship is given by:

$$T_{sec} = \frac{Bf_{bits}}{BU_{bits/sec} - TC_{bits/sec}} \quad (1)$$

where T_{sec} is the time to buffer overflow (seconds), Bf_{bits} is the buffer size (bits), $BU_{bits/sec}$ is the average bandwidth utilization (bits/sec), and $TC_{bits/sec}$ is the maximum tap bandwidth capacity (bits/sec). Note that Equation 1 specifies a linear relationship; however, in practice (i.e., outside a laboratory environment), variations in network packet sizes and rates are typically represented by Gaussian functions. In fact, the

probability of network packet loss in a large network is similar to that in a broadband communications network, for which Gaussian functions have been shown to be very effective [1]. Note that the ability to predict packet loss further supports the task of identifying a potential challenge area that could undermine the foundation testimony (Step 1 in Table 2).

Upon applying Equation 1 to the test case, we confirm the result in Figure 2(b):

$$\begin{aligned} Bf_{bits} &= 1\ MB \times 8\ bits \\ BU_{bits/sec} &= 102\ Mbps \\ TC_{bits/sec} &= 100\ Mbps \\ T_{sec} &= \frac{1\ MB \times 8\ bits}{(102 - 100)\ Mbps} = 4\ sec \end{aligned}$$

The tap manufacturer did not provide information on how the tap implemented queue management; therefore, we assumed that the FIFO algorithm was used. A more thorough analysis might require consideration of alternate solutions. This would be more important in situations where data was gathered under peak conditions, and even more important if an attack on the queuing system could force the tap to drop pre-determined packets.

6. Evaluation of Results

In a hypothetical case involving the collection of forensic evidence using a Net Optics 10/100BaseT Dual Port Aggregator Tap that was subjected to the calibration test described above, one might envision the following dialog between the prosecution's expert witness and the defense attorney regarding the soundness of the evidence gathered by the tap.

Defense Attorney:	Are you confident of the data you collected using this tap?
Expert Witness:	Yes, I am.
Defense Attorney:	Why are you confident?
Expert Witness:	I've tested this tap. I rely on it in the course of business. I understand how it performs, that it drops packets at 4 seconds under maximum load test conditions. Our network was functioning well below capacity during the time in question. I am confident that the Monitor Port captured all the data.

Compare this dialog with what might transpire if the expert witness had relied solely on the vendor's marketing description:

Defense Attorney:	Are you confident of the data you collected using this tap?
Expert Witness:	Yes, I am.
Defense Attorney:	Why are you confident?
Expert Witness:	Well, the manufacturer states the tap has a 100 Mbps capacity.
Defense Attorney:	How do you know this to be true? Have you tested this device?
Expert Witness:	Well, no, I haven't.
Defense Attorney:	Then, how do you know that you've captured all the data during the time in question? Isn't it possible that packets were dropped?
Expert Witness:	Well, I'm certain we captured everything.
Defense Attorney:	But you have no basis to be certain, do you?

The question might arise whether or not the test presented is sufficiently precise to be useful in expert testimony. For example, it does not take into account real-world variability of network traffic, buffer delays or packet collisions. After all, probabilities associated with DNA evidence can be as high as 1 billion to one, perhaps setting high expectations for the precision of other scientific evidence. Nevertheless, it should be noted that it has taken two decades to develop DNA as reliable science. Accepted standards exist for DNA laboratories, for collecting and analyzing evidence, and for training personnel, but these have evolved over time as both the science of DNA and legal case history have evolved [13].

In contrast, network forensic evidence is relatively new and the development of standards is in the earliest stages. Moreover, the acceptability of network forensic evidence has developed differently from DNA. Unlike DNA evidence, where practitioners had to convince the legal system of its validity through a series of court cases, network forensic evidence already is considered admissible [13]. What we anticipate are legal challenges to the credibility of this type of testimony as the legal system gains insight into the technology. We expect these challenges to drive the development of standards, especially as cases that rely on network forensic evidence are won or lost [13].

As standards develop, demands for precision of network forensic evidence will evolve as a function of how crucial the evidence is to making the case and whether or not precision would make a difference. Usually criminal cases are based on a collection of facts, network data representing only one piece of the puzzle. Furthermore, it is often the situation that network data is used to justify a search warrant that produces

additional evidence, which adds to the weight of the case. Given the current state of legal practice, the demand for precision under these circumstances has been low. We believe this situation will change, especially as the defense and prosecutorial bars gain an understanding of network forensics. Consequently, although the calibration test approach presented in this paper is adequate for now, it is only a starting point for developing standards for network forensic evidence.

7. Conclusions

Courtroom admissibility is expected to be a critical requirement for network forensic devices in the future, although it is not so important at the present time [4, 13, 16]. The calibration test methodology described in this paper provides sufficient accuracy for establishing a foundation for legal testimony pertaining to network forensic evidence [13]. Nevertheless, as technology advances and legal case history and courtroom challenges grow, it is expected that the standards of accuracy will evolve and that an eventual calibration standard would include additional considerations of precision.

Our calibration testing work is part of a larger research effort examining network forensic readiness [9, 10]. The idea is to maximize the ability of an environment to collect credible digital evidence while minimizing the cost of incident response [20]. Most approaches to network forensic readiness center on tools and techniques, as opposed to a comprehensive framework for enterprise-wide implementation [7, 14, 21]. To properly embed network forensics, it will be important to determine the standards that will be applied to evidence and, thus, to consider the legal challenges for which a foundation must be laid.

References

[1] R. Addie, M. Zukerman and T. Neame, Broadband traffic modeling: Simple solutions to hard problems, *IEEE Communications*, vol. 36, pp. 2–9, 1998.

[2] Agilent Technologies, Metrology Forum: Basics, Terminology (www.agilent.com/metrology/terminology.shtml).

[3] Agilent Technologies, Metrology Forum: Basics, Why calibrate? (www.agilent.com/metrology/why-cal.shtml).

[4] N. Allen, Are you seeing what you expected? presented at *The Agora*, University of Washington, Seattle, Washington, 2006.

[5] R. Bejtlich, *The Tao Of Network Security Monitoring: Beyond Intrusion Detection*, Addison-Wesley, Boston, Massachusetts, 2005.

[6] S. Bradner and J. McQuaid, RFC 2544 – Benchmarking Methodology for Network Interconnect Devices, IETF Network Working Group (www.faqs.org/rfcs/rfc2544.html), 1999.

[7] B. Carrier and E. Spafford, Getting physical with the digital investigation process, *International Journal of Digital Evidence*, vol. 2(2), 2003.

[8] B. Endicott-Popovsky and B. Chee, NetOptics 10/100BaseT Dual Port Aggregator Tap, Spirent Test Center Technical Report, Advanced Network Computing Laboratory, University of Hawaii at Manoa, Honolulu, Hawaii, 2006.

[9] B. Endicott-Popovsky and D. Frincke, Adding the fourth "R" – A systems approach to solving the hacker's arms race, presented at the *Hawaii International Conference on System Sciences Symposium* (www.itl.nist.gov/iaui/vvrg/hicss39), 2006.

[10] B. Endicott-Popovsky and D. Frincke, Embedding forensic capabilities into networks: Addressing inefficiencies in digital forensics investigations, *Proceedings from the Seventh IEEE Systems, Man and Cybernetics Information Assurance Workshop*, pp. 133–139, 2006.

[11] M. Lawson, Expert Witness Testimony (United States vs. Jimmy Myers Brown (Defendant), Case No. 98-14068-CR, Southern District of Florida, Fort Pierce Division, Fort Pierce, Florida, September 13, 2000), Global CompuSearch, Spokane, Washington, 2006.

[12] D. Nassar, *Network Performance Baselining*, Sams, Indianapolis, Indiana, 2000.

[13] I. Orton, King County (Washington) Prosecutor, personal communication, 2006.

[14] M. Pollitt, Unit Chief FBI CART (Retired), personal communication, 2005.

[15] E. Schultz and R. Shumway, *Incident Response: A Strategic Guide to Handling System and Network Security Breaches*, Sams, Indianapolis, Indiana, 2001.

[16] M. Simon, Chief Technology Officer, Conjungi Corporation, Seattle, Washington, personal communication, 2005.

[17] F. Smith and R. Bace, *A Guide to Forensic Testimony: The Art and Practice of Presenting Testimony as an Expert Technical Witness*, Pearson Education, Boston, Massachusetts, 2003.

[18] P. Sommers, Emerging problems in digital evidence, presented at the *Computer Forensics Workshop*, University of Idaho, Moscow, Idaho, 2002.

[19] Spirent Communications, Spirent TestCenter (www.spirent.com /analysis/technology.cfm?media=7&WS=325&SS=117&wt=2).

[20] J. Tan, Forensic readiness, Technical report, @stake, Cambridge, Massachusetts, 2001.

[21] Y. Tang and T. Daniels, A simple framework for distributed forensics, *Proceedings of the Twenty-Fifth IEEE International Conference on Distributed Computing Systems*, pp. 163–169, 2005.

[22] The Tipmra, The genuine Tipmra speeding ticket defense (www .tipmra.com/new_tipmra/washington_state_speeding_ticket.htm).

[23] U.S. Circuit Court of Appeals (DC Circuit), Frye v. United States, *Federal Reporter*, vol. 293, pp. 1013–1014, 1923.

[24] U.S. Supreme Court, Daubert v. Merrell Dow Pharmaceuticals, Inc., *United States Reports*, vol. 509, pp. 579–601, 1993.

[25] U.S. Supreme Court, Kumho Tire Co. v. Carmichael, *United States Reports*, vol. 526, pp. 137–159, 1999.