The Operational Role of Security Information and Event Management Systems

Sandeep Bhatt, Pratyusa K. Manadhata, and Loai Zomlot | Hewlett-Packard Laboratories

An integral part of enterprise computer security incident response teams, a security operations center (SOC) monitors security incidents in real time. Security incident and event management systems play a critical role in SOCs—collecting, normalizing, storing, and correlating events to identify malicious activities—but face operational challenges.

Computer security incident response teams (CSIRTs) are responsible for receiving, reviewing, and responding to computer security incident reports and activity. Each CSIRT has a defined constituency, such as a corporate, government, or educational organization; a region; or a country. A CSIRT's first task is to monitor security events related to its organization's IT assets. Performing this task is the security operations center (SOC), which is typically a centralized unit in an enterprise IT organization.

Security incident and event management (SIEM) systems are an important tool in SOCs—collecting, normalizing, and analyzing security events from diverse sources—but they must evolve to overcome future scalability issues.

Security Incident and Event Management Systems

A SOC's goal is to monitor security-related events from enterprise IT assets, including the IT network, perimeter defense systems such as firewalls and intrusion prevention devices, application servers, databases, and user accounts. Each asset might be monitored using a variety of sensors and maintain log files of activity. The SOC receives event information from the sensors and log files and triggers alerts indicating possible malicious behavior, both at the perimeter of the network and in the enterprise.

When an alert is triggered, SOC personnel determine whether it was triggered inadvertently—perhaps in response to routine network maintenance—and is harmless, or if the events indicate a strong likelihood of malicious activity. In the latter case, the alert is escalated to a team that coordinates incident response and forensic activities with the owners of the involved servers and applications. In extreme cases, the team must also coordinate with internal human resources, legal and marketing executives, and law enforcement.

A SOC's effectiveness depends on its analytic and forensic capabilities, access to actionable threat intelligence, awareness of the enterprise networks and systems, and internal processes to coordinate responses from organizations across the enterprise.

In the early days, there were relatively few IT security tools, including firewalls for perimeter protection and intrusion detection (IDSs) and antivirus systems (AVSs) for monitoring hosts. Each of these systems came with its own vendor-specific user interface. As

1540-7993/14/$31.00 © 2014 IEEE Copublished by the IEEE Computer and Reliability Societies September/October 2014

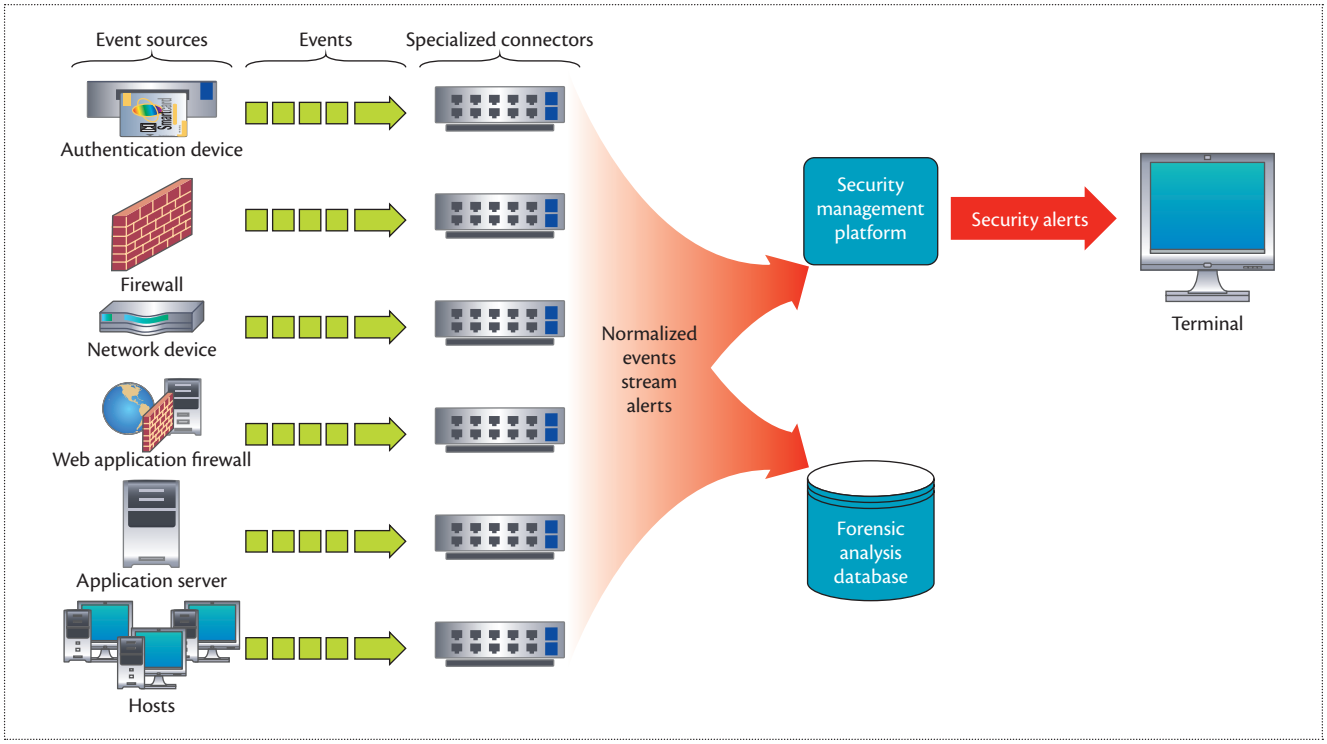

Figure 1. A typical security incident and event management (SIEM) system architecture. The SIEM system accepts inputs from various security devices and sensors. Connectors receive events, parse them, and convert them into a common format.

such tools became more widely used and more tools appeared, two problems arose: first, there were too many user interfaces to manage, and second, there were no tools to correlate events across different security tools. Correlating events across tools became a necessity because individual tools that operate with little or no awareness of the IT architecture trigger too many false-positive alerts, even after careful tuning. On the other hand, when multiple sensors trigger alerts in response to an action, the action is more likely to be malicious.

SIEM systems were designed to meet these challenges: they collect events from diverse sources, each of which might represent events using a vendor-specific schema; normalize these disparate schemata into a common representation; and store these normalized events. Their rule engine triggers alerts from the stored events; the rules allow correlation of events from different sensors. SIEM systems also include auxiliary contextual information, such as up-to-date information on enterprise assets that can be used to write better, context-aware rules and prioritize alerts. SIEM systems' main strength is their ability to cross-correlate logs from diverse sources using common attributes to define meaningful attack patterns and scenarios, which when they occur, can alert security analysts (SAs). Thus, SIEM systems are like radar, detecting objects in a timely manner. Their long-term event retention capability is useful for post hoc forensic analysis as well as investigating and detecting slow and stealthy attacks, including advanced persistent threats (APTs).

Structuring SOCs around SIEM Systems

Figure 1 illustrates an SIEM system's basic architectural components. The SIEM system accepts inputs from various security devices and sensors, including perimeter defense systems (network firewalls and intrusion prevention systems), host sensors (IDSs and AVSs), applications (Web application firewalls and authentication systems), and network sensors. Each device and sensor is configured to output security events with unusual or anomalous behavior that might indicate malicious intent. These events are represented in vendor-, device-, and version-specific schema. So, the SIEM system's first task is to normalize the different representations into a common format to ease further processing and to simplify rule creation and maintenance. As the figure shows, SIEM system connectors—customized for each version, device type, and vendor—receive the events. The connectors parse input events and convert them into a common format, and do so in a scalable manner to keep up with the event source.

Once normalized, events are forwarded to the security management platform and the archival forensic analysis database. The platform maintains and analyzes

a window of events, typically those seen in the past few hours. If necessary, the archival database stores events for a longer term—for instance, three to six months—for further forensic investigation. Although different platforms have different capabilities, a typical archival database can ingest up to 100 Kbytes per second, whereas a typical management platform can ingest approximately 10 Kbytes per second. Specialized hardware can further improve ingestion rates.

The management platform's rule engine applies its rules periodically to events in the window. Whenever a rule triggers a new alert, the alert is sent to the SIEM system terminal for review by a SOC analyst. Each rule captures information about malicious behavior. For example, a rule might look for a large number of failed login attempts within a time window or look for HTTP requests to known malicious websites. The rules are generated from two sources: an analyst can create rules, and the SIEM system can algorithmically generate rules from events, for example, via *pattern mining*—that is, identifying sets of events that occur together frequently within a time window.[1] A rule engine might also use *anomaly detection*—triggering alerts when observed events differ from normal events.

SOC Structure

Modern enterprise SOCs are hierarchically structured around an SIEM system. SOCs typically consist of several levels of SAs—the higher the level, the more experienced the SA and the more specialized his or her function. The lowest-level SAs (usually called level 1) mainly monitor the SIEM system alert screen and triage events, deciding whether they're potential attacks or false alarms. If a rule triggers many false positives, level 1 SAs escalate the rule to a SOC engineer or a higher-level SA for further tuning to reduce false positives. When level 1 SAs can't classify an alert as either an attack or a false positive, they escalate the alert for further investigation.

The large volume of alerts flowing into an enterprise SOC leads to overwhelming amounts of work and long work shifts. Although the number of alerts generated depends heavily on the network environment, SOCs typically aim for 1,000 to 3,000 alerts per day per level 1 SA for manageability. If a network generates more alerts, then SAs sample the alerts. SOC teams usually work around the clock in 8- to 12-hour shifts. This tough schedule and workload generally result in relatively short job retention periods for level 1 SAs, often less than two years including training.

Level 2 SAs are tasked with investigating alerts that level 1 SAs identify as potential attacks. To carry out deeper analysis, level 2 SAs access a wider range of information, including internal sources, such as system logs and asset management systems, and external sources, such as threat activity alerts from public agencies and private corporations. If a breach is revealed, level 2 SAs create a case and forward it to forensics teams and security engineers (levels 3 and above) to determine the attack's extent and impact. If the alert is a false alarm, level 2 analysts work with the security engineering team to fine-tune the rules that created the alert so that it's less likely to trigger false alarms.

An Example SOC

The Hewlett-Packard (HP) enterprise network spans 166 countries and supports more than 300,000 employees. Its Cyber Defense Center (CDC) continuously monitors the network—HP's version of a state-of-the-art SOC—and is staffed with level 1 and 2 analysts, who are supported by dedicated forensics analysts. The CDC's structure and its deployment of SIEM systems are representative of SOC deployments for large, global enterprises and government organizations.

We estimate that HP's enterprise network generates 100 billion to 1 trillion security events each day. Collecting, storing, and analyzing all these events is nearly impossible, so the CDC focuses on several important event sources such as HTTP proxy, DNS, and antivirus logs. Its elaborate SIEM system infrastructure involves hundreds of load-balanced connectors, more than 100 instances of archival databases, and multiple instances of the SIEM system manager. The infrastructure currently processes approximately 3 billion events per day; this could grow to 30 billion events in the near future.

Cost-Benefit Analysis

SIEM systems' popularity in SOCs is due primarily to their ability to handle a large number of events from many different sources. When enterprise networks were smaller and generated fewer events, SIEM systems weren't very popular among enterprise network administrators. As enterprise networks started growing due to the addition of new devices, applications, and employees, the number of events generated also grew. In addition, network administrators started collecting events from more sources. Hence, over the past decade, SIEM systems have become an indispensable tool for handling enterprise security events. SIEM systems are arguably the most important tools in SOCs today, and we expect the trend to continue.

However, operating large-scale SIEM systems requires a large budget. A typical management platform might cost US$80,000, and an archival database might cost $20,000. Hence, a large SIEM system with hundreds of connectors, a few hundred archival databases, and multiple platforms might require $3 to $5 million in up-front hardware cost and additional

yearly maintenance cost. If a SOC staffs 20 SAs, the yearly operating cost might be upwards of $5 million.

Enterprise-scale SIEM systems need significant investment in both hardware and manpower, and SIEM systems and SOCs must continue to deliver to justify the investment.

Operational Challenges

SOCs confront various operational challenges when using SIEM systems, driven primarily by the scale and complexity of the enterprise being monitored and the rate at which events arrive from security devices and sensors.

Rule Creation and Management

Having all the network and host logs at the SAs' fingertips is attractive because the more information a SOC has, the better its situational awareness. However, this comes at the cost of transforming an SIEM system's data management to big data management, which turns storage, search, sharing, transfer, analysis, and visualization into challenges. One aspect of this problem is the system's inability to efficiently execute complex queries, severely limiting SAs' ability to write c omplex correlation rules. A more problematic aspect is the number of false alarms that the SIEM system rules tend to trigger. Because benign events outnumber malicious ones, even a low false-positive rate will produce many false alarms,[2] which the SOC might not have the capacity to deal with. Therefore, SIEM system rules must have extremely low false-positive rates to be usable in practice.

> **Communicating the right kind and amount of information between enterprise operations and the SOC in an automated way is essential.**

Most of the time, SIEM system analysts need to write very specific rules to capture an attack, but this means the system might miss other forms of that attack. Thus, there's always a tradeoff between false-positive and false-negative rates. To prevent false negatives—that is, detection misses from overly specific attack rules—engineers resort to generic rules, so that an activity with even a remote possibility of indicating an attack will trigger an alert. Then, analysts are responsible for monitoring the SIEM system to distinguish the true alarms from the enormous number of false ones. Many SOC teams have limited resources to process overwhelming volumes of events. Thus, the SOC enters a vicious cycle of accumulating more and more alerts that SAs must process each hour.

In talking with many SOC teams, we found it's acceptable to triage an event in 10 minutes; some teams would like to reduce this to one minute or less! Such severe time restrictions force SAs to sample alerts from the events list. Although the number of alerts on SA screens is proportional to the size of the logs flowing into the SIEM system, some professionals claim that by writing the right SIEM system rules and applying the right management techniques, they can dampen the relation between the volume of logs and alerts. This remains speculative; with the explosive growth in data rates, it's difficult to see how SIEM system processing rates can keep up under cost constraints.

Lack of Contextual Information

Another challenge SOCs face is isolation from enterprise network operations. SOC personnel aren't involved in the details of configuring, testing, and maintaining enterprise assets. Routine activities such as patching, backup, and testing might trigger alerts in SIEM systems designed to detect security breaches, and tracking down the cause of such alerts creates unnecessary overhead.

In an interview, a senior SA pointed out the importance of automatically collecting detailed host configurations, servers, devices, and user information. In principle, this information can be correlated with SIEM system alerts to significantly reduce false-alarm rates. However, collecting and maintaining this information, especially in large networks, is challenging. Instead, such information is often communicated in an informal, even ad hoc manner, either verbally or via email. In one incident, an SA had to contact the network operations team about potential malicious activity in the internal network, which turned out to be a spurt in traffic from a patching server. The SOC saw probing alerts on its screens; these were manually tracked to the patching server and eventually declared false positives.

Information communicated informally usually falls through the cracks when SOC analysts change shifts, thus exacerbating the problem. Instead of storing crucial contextual information in SIEM systems, all too often, SOCs rely on SAs to maintain this information. Unfortunately, this information is lost when SAs leave and replacements are hired.

In general, although isolating a SOC from the enterprise systems' routine maintenance activities is a reasonable objective, communicating the right kind and amount of information between enterprise operations and the SOC in an automated way is essential to reduce the SOC's load and to achieve more effective security monitoring.

Ad Hoc Use of Long-Term Data

Current SOC operations use SIEM system features in an unbalanced manner. Most teams focus on short-term alerting functionality and ignore the long-term retention feature. SOCs usually monitor a rolling and narrow time window of events—typically an hour or two. This limits their ability to detect stealthy, slow-advancing attacks, especially APTs. In our interviews with SOC analysts, we found that many teams were aware of this issue and tried to mitigate it by sampling the raw logs randomly with the hope of finding attack patterns. SOC analysts should give more attention to the SIEM system's retention feature and develop analytic solutions that help in visually revealing the patterns of the slow and stealthy attacks. Running these analytics also helps to guide the SIEM system rule-writing process by unveiling unknown attacks. This practice shifts the SOC team from a reactive to a proactive state.

Technical Challenges and Opportunities

We believe that SIEM systems will continue to be an integral part of SOCs; however, they must evolve to deal with event collection, storage, analysis, and visualization challenges. The key challenge SIEM systems face is the scale of events enterprise networks generate from hardware devices, software applications, and actions of people in the network. More complex applications and devices tend to generate more events; for example, accessing a modern webpage might generate several HTTP requests for content embedded in the webpage. Therefore, an enterprise's number of events is proportional to the number of employees, the number of network devices and applications, and the devices' and applications' complexity. The number of events will grow as enterprise administrators enable event logging in more devices and applications; the number will also grow each year as corporations hire more employees and add new and complex devices and software applications to their network, for instance, bring-your-own-device scenarios.

Event Collection

Collecting events in a scalable manner is a challenge. For example, if we decide to collect 1 trillion events per day, our SIEM system must support an ingestion rate of 12 million events per second. The required rate is orders of magnitude greater than the state of the art; supporting this rate requires significant investment, such as additional connectors. SIEM systems also face logistical challenges in data collection—an event source of interest might generate too few events or too many benign events that are of no use to SOC operations. For example, DNS servers often log only DNS queries and ignore DNS responses. Similarly, laptop and desktop syslogs might contain too many benign events. Hence, SIEM systems might need their own event generation and filtering solutions to cope with problems of scale.

SIEM systems also face input validation challenges. As we collect events from many sources, the data becomes increasingly noisy. For example, an event source might be unavailable or might not be able to populate all fields in an event. The data might also contain malicious events. If attackers compromise a data source or a communication channel, they might be able to inject malicious events into SIEM systems. Hence, SIEM systems must evolve to prevent attackers from generating malicious events and detect and filter out malicious and noisy events.

Ultimately, the challenge is to determine the "right" events to collect. Given a security problem to solve, collecting and correlating the right events is much easier than dealing with all enterprise events. But how do we determine the right events a priori? Are there problems for which event sampling is sufficient? There will always be the fear of missing important events if not everything is collected.

Event Storage

SIEM systems must balance storage costs with analysis requirements. For example, if one event requires approximately 400 bytes of storage and we achieve 10x data compression, then 1 trillion events per day will require 40 Tbytes of storage. Although storing the data for perpetuity would be ideal for analysis and forensics, storage costs make this impractical. Thus, we need to define a data retention period that makes an optimal tradeoff between storage costs and analysis requirements. In addition, regulatory compliance might require sensitive events to be deleted after a time period.

SIEM systems also face a tradeoff in storage architecture, as forensic analysis and real-time analysis have different characteristics. In forensic analysis, we make infrequent queries on large data volumes; thus, write-optimized databases with high ingestion rates might be suitable. As long as SIEM systems can store events at a high rate, a reasonably higher query response time or a small delay in data availability is tolerable. However, read-optimized databases might work better for real-time analysis. We might prefer the ability to read stored data quickly for analysis, even though we have to tolerate a low ingestion rate or a small delay in data availability.

Finally, the large event dataset might contain private and sensitive information such as employee Web browsing history. Hence, SIEM systems face the challenge of securing stored events from unauthorized access.

Event Correlation and Analysis

SIEM systems correlate events across multiple event streams and look for known event patterns to identify

attacks and other security-relevant events. For example, they might correlate HTTP proxy and antivirus product logs to detect malware downloaded to endpoint devices. However, performing correlations and identifying patterns at the scale of large enterprises remain challenges. Moreover, SIEM systems must perform more sophisticated analysis to derive true value from the collected events.

Scalable analysis algorithms that handle 1 trillion or more events per day face significant challenges. First, identifying attacks from event streams is more art than science—no definition of attacks exists, so SOC analysts use heuristics derived from past experience to identify attacks and other relevant events. Analysis algorithms should automatically identify patterns of interest from large event streams, but automating a heuristics-driven process is difficult.

Second, even if an algorithm can identify attacks today, it might not work tomorrow as adversaries adapt, enterprise networks change, and employees' behaviors change. Hence, the algorithm must learn and evolve continuously.

Third, the problem of false positives becomes more acute as SIEM systems collect more data. Because benign events outnumber malicious events by orders of magnitude, an extremely low false-positive rate might still produce too many false positives to be usable in practice. Hence, analysis algorithms might not be able to make a scalability–accuracy tradeoff.

Fourth, even if an analysis algorithm produces no false positives, it might produce more true positives than SOC analysts can handle. Hence, SIEM systems will have to prioritize the true positives for SOC analyst consumption.

Fifth, more events might lead to statistically significant but ultimately meaningless correlations.[3] When dealing with high-dimensional data, many unrelated variables can have high correlations, which will manifest as false positives. Hence, analysis approaches should be able to filter out spurious correlations.

Finally, the hardest challenge is inferring human intent from machine logs—analysis algorithms will have to infer attacker and user intent from event streams to identify true attacks, as both malicious attacker actions and benign user actions might generate the same event patterns.

Visualization

We believe that SIEM systems will never reach the maturity level needed to replace human analysts in SOCs. At best, they'll be tools in analysts' and network administrators' decision-making processes. Hence, SIEM systems face the challenge of summarizing analysis results and presenting them so that humans can make more effective and efficient decisions, for instance, identifying a new attack or deciding which security alerts to respond to. SIEM systems must develop visualization techniques that aid humans in gathering information from large quantities of data, provide context information in a timely manner, and work at different organizational levels, such as system administrator and higher-level management.

Toward Addressing the Challenges

Multiple SIEM system vendors have offered different approaches to improve SIEM systems' capabilities to collect, store, and correlate events in large enterprise networks; however, progress is necessary to address scalability issues. For example, because complex correlation is time consuming, analysts typically avoid creating feature-rich correlation rules that incorporate many information sources to capture sophisticated and stealthy attacks.

To the best of our knowledge, no research directly addresses SIEM system challenges. However, we believe that advances in many fields of computer science will significantly impact SIEM systems. For example, advances in storage systems—especially nonvolatile memory—will help with storing more event data at lower cost. Similarly, advances in parallel and distributed computing, especially in big data analysis, will provide the platform for scalable analysis. For example, a distributed correlation engine might handle more complex rules than traditional SIEM systems. There's also recent work on using big data analysis to identify actionable security information from very large event datasets. Ting-Fang Yen and her colleagues analyze HTTP proxy logs to identify suspicious host activities—they extract features from logs, then use clustering to find outlying suspicious activities.[4]

There's a long line of research on alert correlation as a way to increase the features available to make decisions, building on the assumption of the impracticality of achieving meaningful results on the basis of a single event such as a network packet.[2,5,6] However, alert correlation solutions tend to have false correlations from the large amount of low-quality events that SIEM systems handle. This has led to research on alert prioritization—that is, identifying higher-quality alerts that analysts should focus on. Researchers introduced multiple alert prioritization approaches, some using probability theory and Dempster-Shafer theory.[7,8]

Big data visualization is a very active research area.[9] Data visualization specifically for security has also been explored.[10] Advances in these two areas will help address the big data visualization problem that SIEM systems face.

Although SIEM systems provide a solid technical foundation for SOCs, we believe further advances are needed to create a more adaptive, context-aware, flexible, holistic, and social solution. This solution should capture organization-specific knowledge as feedback from the SOC to guide the investigation processes and reduce the volume of events. The goals should be to

- give the SOC the freedom to design and adapt its processes per incident,
- manage incidents by collecting all related information and communications in one place and help the system learn from previous incidents (to reduce repetitive investigations triggered by similar false alarms, for example), and
- automate and customize the repetitive aspects of each role in the SOC team by presenting contextual information to SAs along with the alert (as opposed to the current practice of having SAs manually ferret out relevant information from multiple sources). ■

References

1. R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proc. ACM SIGMOD Int'l Conf. Management of Data* (SIGMOD 93), 1993, pp. 207–216.
2. S. Axelsson, "The Base-Rate Fallacy and the Difficulty of Intrusion Detection," *ACM Trans. Information System Security*, 2000, pp. 186–205.
3. O. Ogas, "Beware the Big Errors of 'Big Data,'" *Wired*, 2013; www.wired.com/opinion/2013/02/big-data-means-big-errors-people.
4. T.-F. Yen et al., "Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks," *Proc. 29th Ann. Computer Security Applications Conference* (ACSAC 13), 2013, pp. 199–208.
5. F. Cuppens and A. Miege, "Alert Correlation in a Cooperative Intrusion Detection Framework," *Proc. IEEE Symp. Security and Privacy*, 2002, pp. 202–215.
6. B. Morin et al., "A Logic-Based Model to Support Alert Correlation in Intrusion Detection," *Information Fusion*, 2009, pp. 285–299.
7. L. Zomlot et al., "Prioritizing Intrusion Analysis Using Dempster-Shafer Theory," *Proc. 4th ACM Workshop Artificial Intelligence and Security* (AISec 11), 2011, pp. 59–70.
8. Y. Zhai et al., "Reasoning about Complementary Intrusion Evidence," *Proc. 20th Ann. Computer Security Applications Conf.* (ACSAC 04), 2004, pp. 39–48.
9. S. Liu et al., "A Survey on Information Visualization: Recent Advances and Challenges," *The Visual Computer*, Springer, 2014, pp. 1–21.
10. R. Marty, *Applied Security Visualization*, Addison-Wesley, 2009.

Sandeep Bhatt is a researcher at Hewlett-Packard Laboratories. His research interests include network data security analysis, managing end-to-end access control in enterprise networks, and algorithms for parallel computation, network communication, and VLSI layout. Bhatt received a PhD in computer science from MIT. Contact him at sandeep.bhatt@hp.com.

Pratyusa K. Manadhata is a researcher at Hewlett-Packard Laboratories. His research interests include security and privacy, with an emphasis on big data analysis for security. Manadhata received a PhD in computer science from Carnegie Mellon University. He's a member of the ACM, IEEE, and Usenix. Contact him at manadhata@hp.com.

Loai Zomlot is a postdoctoral research associate at Hewlett-Packard Laboratories. His research interests include cybersecurity, with a special focus on intrusion detection and analysis. Zomlot received a PhD in computer science from Kansas State University. Contact him at loai.zomlot@hp.com.