Lessons Learned through Customer Discovery in a Provenance-based Security Start-Up

Akul Goyal
Provenance Security, Inc.
akulgoyal@pravsec.com

Adam Bates
Provenance Security, Inc.
adambates@pravsec.com

Abstract*—Provenance-based security applications are showing tremendous promise in the academic literature, but successfully transition these technologies to practice will require community stakeholders to demonstrate the business potential of provenance analysis.** ***Customer Discovery **is a structured process through which early-stage start-ups can validate the commercial potential of an idea through direct interaction with potential customers. As a provenance-based security start-up, we conducted hundreds of customer discovery interviews, and believe that many of our findings would be of interest to the broader academic community. In this position paper, we summarize our findings and consider how they could inform future research on provenance analysis.**

I. Introduction

Provenance analysis research demonstrates extraordinary potential across security applications, yielding breakthroughs in threat detection (e.g., [4, 10, 15, 23]), investigation (e.g., [1, 8, 16, 22]), and resolving existing product shortcomings (e.g., [6, 11, 12, 14, 19, 21]). In fact, many of these works have proven their systems in contolled live environments - showing the potential for provenance. However, publications alone remain insufficient to drive commercial product evolution. Provenance is a paradigm shift that potentially requires re-architecting existing security infrastructure, a financial risk that adopters must weigh against limited evidence of real-world efficacy. Successfully transitioning provenance-based security applications to practice requires start-ups to de-risk the technology and validate underlying business hypotheses.

Customer Discovery, a standard start-up procedure, tests whether products or ideas solve genuine problems for actual customer groups [2]. The process involves forming hypotheses about potential business ideas, testing them through customer interaction, then iteratively refining or pivoting based on findings. Customer discovery determines problem-solution fit before significant product development investment, while providing academic scientists and technologists early exposure to business contexts. The National Science Foundation's Innovation Corps (I-Corps) customer discovery training is a seven week program where academics interested in commercializing their research are introduced to formal frameworks they can use to conduct interviews and extract insights.

Founded by academic researchers, our cybersecurity startup participated in regional and national NSF I-Corps programs to engage with stakeholders throughout the threat detection and response ecosystem. To ensure a comprehensive perspective, we interviewed professionals across a spectrum of job roles—including Chief Information Security Officers (CISOs), Security Operations Center (SOC) managers, cyber analysts, and security engineers—spanning diverse sectors such as finance, healthcare, technology, and government. The majority of these dialogues took place at key industry conferences, including Splunk's developer conference, BSides Chicago, BSides SF, and FIRST. Beyond potential clients, we also engaged with current security providers to assess the capabilities of existing solutions. This multi-faceted approach enabled us to synthesize a robust understanding of current practices, incorporating insights from both frontline practitioners and the vendors serving them.

We believe that our customer discovery findings are of value to the community of provenance researchers. Most findings apply broadly to provenance-based security tools beyond our specific product development. We confirmed with actual stakeholders many industry reports motivating academic provenance research (e.g., [3, 5, 7, 13]), while providing nuanced understanding of how these phenomena manifest in practice. Many findings illuminate future directions for fundamental scientific research on provenance security.

This position paper summarizes high-level findings from our customer discovery process. These interviews constituted business activities conducted using NSF I-Corps program procedures, not human subjects research. We did not consult an Institutional Review Board; as a start-up rather than university, our organization operates outside National Research Act jurisdiction. This work intends to provoke thought and discussion on commercial provenance technology transfer potential, but should be viewed as anecdotal rather than hard scientific data.

II. Lessons Learned

A. Lesson #1: The Problems Are Real (Kind Of)

Our initial customer discovery interviews sought to identify whether problems motivating current academic research align with challenges security practitioners face in practice.

Workshop on Attack Provenance, Reasoning, and Investigation for Security in the Monitored Environment (PRISM) 2026
23 - 27 February 2026 , San Diego, CA, USA
ISBN 978-1-970672-00-8
https://dx.doi.org/10.14722/prism.2026.23011
www.ndss-symposium.org

Understanding this alignment required mapping analysts' day-to-day workflows to establish how various challenges impact operational effectiveness. We began by asking security analysts to describe their routine processes, identifying which tasks consumed the most time and which obstacles impaired their investigative capabilities. From our investigations, we began to learn how the problems that academia used to motivate their work were reflected by security practitioners.

A common motivation for provenance research [11, 12] is alert fatigue: the substantial burden false positive alerts practitioners face daily. We anticipated analysts would spend considerable time on alert triage, trying to distinguish genuine threats. While some practitioners reported high alert volumes, more often we encountered the opposite. Many practitioners reported minimal alert fatigue, handling only $5 - 10$ alerts weekly. As one medical startup security engineer stated, "if a security team is dealing with more than 1 alert per 1000 devices, they have not correctly set up their environment." Enterprise detection relies predominantly on signature-based rules created by vendors and MDR services. These providers distribute pre-defined rules to customers, but critically, the rule logic remains opaque—organizations typically only enable or disable rules, not inspect or modify them. Facing high false positive rates from generic rules, organizations routinely disable noisy detections through "rule engineering", effectively reducing alert volumes by sacrificing visibility into potential attack behaviors. Smaller teams without dedicated security analysts outsource alert management to Managed Service Providers (MSP) who perform this pruning. These mechanisms cultivate a belief that alert fatigue indicates poor SOC configuration rather than systemic detection challenges. In reality, the problem trades volume for visibility, as disabled rules create blind spots. Notably, well-resourced teams that have a higher tolerance for false alerts, sometimes deliberately retain noisy rules when the potential impact of missing specific attacks justifies the false positives.

Another motivating problem for provenance research is the lengthy investigation times associated with resolving alerts. Academic literature [17, 18] often assumes analysts trace backward from alerts to reconstruct attack sequences. However, our interviews revealed investigations are typically brief, with extended resolution times reflecting human coordination delays rather than analytical inefficiency. As one airline security analyst explained, "The alerts that take the longest to close require us contacting the device owner and getting a response back from them." SOCs structure incident response into three tiers, with alerts escalating as they demand deeper analysis. At each tier, investigations follow increasingly deeper playbooks: prescribed query sequences that validate or dismiss alerts. Comprehensive threat hunting—tracing back to identify all processes and files associated with an alert—remains uncommon within industry. As a Fortune 500 threat analyst stated, "Active incident response is something that most companies do not have the time to do." Often analysts will face service licensing agreements (SLA) which dictate how fast they must close alerts preventing non-scripted investigations. Moreover, playbooks ensure consistent investigations where query based backtracking can deviate based on the analyst's experience.

Finally, provenance research often leverages Advanced Persistent Threat (APT) examples [4, 10, 22] to demonstrate the necessity of causal analysis. However, the majority of security don't prioritize defense against such sophisticated campaigns. Instead, these teams allocate most of their bandwidth to mitigating high-velocity threats—such as ransomware and DDoS—which constitute the bulk of actionable incidents. Consequently, many practitioners do not view APTs as a primary operational concern. As a Midwest security engineer noted, an APT requires "the attacker to execute several attack steps"—intermediate actions that they believe should be detectable through conventional means. This perspective is further complicated by organizational scale. Teams at smaller organizations often questioned whether they were targeted by sophisticated actors given their size. Conversely, larger organizations with mature security structures still struggle to dedicate resources toward hunting APTs due to the sheer vastness of their infrastructure. As the head of detection at a Fortune 500 company observed, "almost every large company has an attacker inside their system that they do not know about."

Overall, we found that these while these motivating scenarios existed in practice, they were either not perceived in the same form that academic work assumes. Instead, changing how these problems were represented helped resonate with stakeholders when discussing the capabilities of provenance analysis.

B. Lesson #2: Provenance faces a Knowledge Translation Gap

The second challenge we faced in customer discovery process was one of knowledge translation: security analysts consistently misunderstand how provenance analysis worked and what it could provide. We attribute these misunderstandings to various factors that shape how practitioners conceptualize security. First, many traditional security logs lack the causal relationships necessary for provenance analysis, fundamentally limiting what current tools represent. Second, the industry's entrenched tabular data[1] architecture—optimized over decades for correlation-based detection—enforces thinking about security logs as independent events. These structural limitations create a disconnect where analysts believe they possess provenance capabilities through existing correlation rules and workflow automation, even when product documentation confirms no such capabilities exist. This market misconception presents a significant adoption challenge for provenance technologies.

1) Analysts Reason Differently About Security Data: Security analysts operate within an ecosystem fundamentally designed around correlation rather than causation, shaping their mental models of threat analysis. A security rule engineer at a major technology company exemplified this mindset, stating he would adopt causal analysis only if he could "see the

[1]We use the term "tabular" to refer to both true tabular data as well as temporally-ordered semi-structured event data, which is the dominant paradigm in commercial products.

underlying correlational rules that [causally] relate log events," revealing how correlation and causality are conceptually conflated throughout the industry. This reasoning pattern stems from the structural characteristics of enterprise security data. Most security logs – authentication records, network flows, application events – are logged as independent events where causal relationships remain opaque without broader system context. IAM logs, for instance, record user authentication attempts across machines but provide no mechanism to establish causal dependencies between login sequences. Analyst training therefore emphasizes manually piecing together disparate log sources to investigate potential attack behaviors. As previously stated in Section II-A, Tier-1 analysts operate from standardized playbooks—structured query sequences designed to validate whether alerts represent genuine malicious activity. This playbook-driven approach reinforces correlation-based investigation patterns where analysts answer predetermined questions across multiple log systems to decrease response times and standardize analysis rather than trying to recreate the series of log events that lead to the alert occurring. This correlation-centric mindset creates significant adoption barriers, as provenance requires analysts to reconceptualize their investigative workflows around causal reasoning rather than correlation matching.

2) "Yeah, we've got that.": This knowledge gap is further exacerbated by current security tooling, which has developed features and products that superficially resemble provenance capabilities without delivering causal analysis. Existing products, from collection agents to SIEMs to data lakes, are architected on the assumption that security events are independent, discrete records optimized for time-ordered storage and retrieval. This tabular model aligns with established storage paradigms, simplifies query operations, and reinforces analysts' existing mental models. Efforts to connect and relate events within these systems emerge as architectural afterthoughts, delivering diluted approximations of provenance capabilities. This architectural entrenchment creates significant market confusion when introducing provenance-based approaches. Security teams consistently report they already possess provenance capabilities while fundamentally lacking needed relationships. To our knowledge, no commercial security tool currently offers actual provenance capabilities derived from information flow analysis. Below, we examine some of the existing tools mis-identified as provenance analysis tools.

User-Entity Based Alerting: When discussing provenance-based alert grouping—where alerts are clustered according to dependency relationships—security analysts consistently referenced User and Entity Behavior Analytics (UEBA) as providing equivalent capabilities. A CISO and security engineering team at a mid-sized Chicago MSP exemplified this conflation, explaining they "automated backtracing from an alert by using related fields...establishing causality" through Elasticsearch's UEBA implementation. UEBA, pioneered by Splunk and now ubiquitous across SIEM platforms, correlates alerts by matching common attributes (users, devices, processes) within defined time windows. Security engineers write correlation rules leveraging UEBA to group alerts based on known attack patterns, then triage incidents according to emerging grouping patterns. However, UEBA's reliance on coarse-grained features produces two critical failure modes: alert clusters overwhelming analysts with loosely related events, or fragmented groupings revealing only partial attack campaigns spanning multiple entities. Even when UEBA grouping succeeds, analysts must still conduct manual query-based investigations to establish root causes and attack impact. Despite these limitations, UEBA's intuitive correlation rules enable analysts to quickly hypothesize likely attack scenarios. UEBA is also able to correctly group attack indicators in less complex scenarios where the attack does not span multiple entities (i.e., one endpoint/ueser). Transitioning security analysts to provenance-based approaches requires demonstrating UEBA's limitations and how dependency-based analysis provides more accurate, complete attack reconstruction.

Process Trees: When shown provenance graph examples during customer discovery calls, security analysts consistently referenced graph features in commercially available equivalents. But they were *always* process trees. Process trees represent parent-child relationships within EDR log data. Analysts interpreted the causal dependencies visualized in provenance graphs as equivalent to these process trees, asserting that process execution dependencies were actively used in current investigation processes. A security engineer at an R1 University exemplified this perspective, explaining he "would use the process tree to triage the severity of the alert," escalating investigations when suspicious processes appeared in the tree, otherwise marking alerts as false positives. Indeed, process trees are useful in that they capture a vital causal relationship, process lineage. The difference is that provenance graphs include process lineage alongside other essential causal relationships like data flow and interprocess communication. Nonetheless, analysts perceived provenance as replicating familiar capabilities rather than introducing novel analysis. Provenance graphs' distinguishing advantage – identifying dependencies beyond process execution chains such as process-file-process interactions – was not perceived as a benefit, as such attack techniques are rarely investigated by analysts as discussed in Section II-A.

Artificial Intelligence: When discussing provenance graphs' potential to improve investigation efficiency, security analysts frequently questioned how they compare to artificial intelligence solutions. The rise of generative AI and agent-based analysis has (allegedly) enabled security products to automate substantial investigation work, including executing the standardized playbooks analysts currently perform manually. This automation within familiar formats creates significant competitive pressure against provenance adoption. The head of investigations at a medium-sized technology company explained his team used AI to analyze each alert and "create a story of what the attacker was doing in their system." However, among other potential problems early adopters may face in the future, LLM-based tools face a critical barrier: output inconsistency. Hallucinations in cybersecurity contexts can have

cascading organizational impacts based on analyst decisions. Consequently, security analysts remain wary of automation that could produce lasting organizational harm. Provenance's ability to deterministically relate alerts offers a distinct advantage over probabilistic AI approaches. Nevertheless, the specialized knowledge required to interpret provenance graphs creates adoption barriers that AI tools operating within analysts' existing conceptual frameworks do not encounter.

C. Lesson #3: Deploying Provenance Commercially

We now shift focus from understanding the gap between academic research and commercial perception to examining deployment requirements for provenance technology in commercial settings. This reflects the next phase of our I-Corps customer discovery process, where we assessed the current commercial security ecosystem and identified where provenance-based solutions integrate within existing architectures. Customer interviews revealed two fundamental constraints governing commercial viability: infrastructure impact (the extent of deployment changes required) and operational criticality (whether the solution affects security-critical functions). Successfully transitioning provenance research to commercial products requires explicitly addressing these constraints while maintaining clear value justification, ensuring demonstrable benefits warrant the operational costs of provenance-based technologies.

1) Problems with logging: Successfully translating provenance to commercial settings requires understanding the data environment that provenance will analyze. Provenance analysis fundamentally depends on comprehensive system logging, which could theoretically be achieved through custom collection agents designed specifically for provenance tracking. However, security organizations already operate established logging infrastructure, making deployment of new collection agents an unacceptable infrastructure disruption with significant security risk. This reality forces provenance technologies to integrate with existing security telemetry rather than replace it, introducing challenges inherent to analyzing current enterprise logging ecosystems.

EDR Log Heterogeneity and Incompleteness: EDR logs represent the closest commercial analogue to the system-level audit data that academic provenance research assumes, making EDR integration the natural entry point for deployment. However, two fundamental obstacles complicate this integration: vendor-specific format variations and systematic omission of events necessary for dependency reconstruction. EDR vendors capture fundamentally different data, preventing uniform analysis approaches. Carbon Black represents process-to-file interactions as `endpoint.event.filemod` events with explicit `ACTION_FILE_OPEN_WRITE` access modes, while SentinelOne and Microsoft Defender follow similar patterns under vendor-specific field names. CrowdStrike diverges entirely, employing 58 specialized event types like `ELFFileWritten` and `PDFFileWritten` where access patterns are encoded within event type rather than separate access fields. Provenance reconstruction requires translating these disparate representations into unified schemas. Given the proliferation of EDR platforms across enterprise environments, such translation must be automated to achieve practical scalability. Beyond schema heterogeneity, all EDR agents systematically exclude certain events to maintain computational and storage efficiency. Academic provenance research typically assumes entity-level tracking—already an abstraction from the byte-level precision required for complete information flow analysis. Commercial EDR systems operate at even coarser granularity, omitting computationally expensive operations like individual read/write events. This systematic absence of fine-grained operations undermines dependency identification, which relies on such events to establish causal relationships between system entities. Translating provenance to practice thus requires accommodating substantial gaps in event coverage while maintaining analytical utility—constraints largely absent from academic threat models.

The Multi-Source Log Integration: While practitioners view EDR logs as the highest-quality telemetry, offering rich contextual information in structured formats with manageable noise, EDR simultaneously represents the most incomplete data source in enterprise environments. Agent deployment requires permissions and system access frequently outside security teams' control, creating coverage gaps fundamentally incompatible with academic assumptions of universal instrumentation. This incompleteness prohibits exclusive reliance on EDR for investigation workflows. Instead, practitioners routinely correlate authentication records, network captures, cloud activity logs, and application events to reconstruct attacker progression through enterprise infrastructure. Customer conversations about provenance consequently emphasized cross-stack integration over EDR-only approaches. A healthcare CISO articulated this directly: "The holy grail of security investigations is looking at a single pane of glass." Confining provenance to EDR telemetry creates exploitable visibility gaps where uninstrumented devices enable detection evasion. Practical deployment therefore demands reconstructing information flow relationships across heterogeneous log sources, including scenarios where EDR coverage is sparse or absent. Successfully unifying diverse telemetry streams into coherent provenance representations addresses a core operational need—delivering value that enables commercial adoption.

2) Existing "Tabular" Representation: Provenance analysis operates on graph-structured dependency relationships, yet enterprise security infrastructure universally stores telemetry as time series of semi-structured data, optimizing for temporal querying and correlation. Customer discussions consistently surfaced concerns about the computational and storage costs associated with provenance deployment. A security engineer at a technology company articulated this skepticism directly: "It is hard to justify the overhead of provenance-based analysis if similar results can be achieved with creating correlational rules that group alerts together." Academic research has traditionally assumed access to pre-constructed provenance graphs ontop of which it can conduct analysis. However, materializing complete provenance graphs from tabular data would double

storage requirements and demand substantial computational resources for organizations managing petabyte-scale log retention. Moreover, migrating to graph-based storage infrastructure would constitute precisely the disruptive architectural overhaul that commercial customers reject. Provenance analysis must therefore operate directly on existing tabular representations. Bridging the architectural gap between graph-based algorithms introduced in academia to table-oriented enterprise platforms represents a critical translation challenge for practical adoption of provenance.

D. Lesson #4: The Provenance User Experience

As customer discovery interviews progressed, we developed a product demo targeting provenance-based threat investigation to understand how security analysts could integrate provenance into their workflows. The demonstration presented an interactive system containing data from a laboratory simulation spanning several weeks, capturing both benign user behavior and an APT-style attack across multiple machines. A dashboard summarized simulation events into incidents—groupings of alerts based on causal relationships—with each incident accompanied by a provenance graph visualizing the causal interactions between constituent raw events. While we applied minimal edits to enhance readability, we preserved comprehensive detail to support thorough investigation. This approach enabled direct observation of practitioner reactions to provenance visualization, revealing critical insights about fundamental usability challenges that extend beyond technical capability to interface design and cognitive fit with existing analytical practices.

1) Visualizing Provenance Graphs Is Hard: Practitioner feedback consistently centered on interpretability challenges when interacting with our provenance visualizations. Two specific design tensions emerged: the spatial organization of nodes and the appropriate level of detail for effective analysis.

Node Placement: Security engineers reported that our initial visualizations were too complex to parse effectively. They expected clear left-to-right chronological ordering to track threat actor progression through their systems—a mental model aligned with temporal narratives of attack sequences. However, maintaining strict chronological ordering introduces significant layout complexity. Provenance graphs inherently preserve temporal structure through versioning, which represents system entity state changes according to information flow. Versioning generates multiple nodes per entity as its state evolves, exponentially increasing the number of visual elements beyond non-versioned representations. Moreover, positioning nodes to enforce exact left-to-right temporal ordering produces extraordinarily wide visualizations that analysts struggle to traverse, forcing horizontal scrolling that disrupts cognitive continuity and obscures attack progression.

Node placement for provenance graph readability remains largely unexplored in academic research. While general graph layout algorithms exist [9, 20], these approaches lack the temporal ordering constraints and causal semantics inherent to provenance visualization. Determining optimal layout strategies that balance temporal clarity with spatial efficiency represents a critical research gap for creating analyst-friendly explanations.

Appropriate Level of Detail: The head of investigation at an R1 university observed that junior analysts lack the experience to interpret the low-level system events our visualizations captured. Because our graphs displayed raw events between alerts—including granular system execution calls—they provided exhaustive detail that most SOCs do not utilize during triage and investigation. Security teams prioritize high-level attack progression patterns that quickly reveal adversary actions and anticipated next moves rather than comprehensive system-level traces. This granularity overwhelmed analysts, effectively preventing practical use. Practitioners requested highlighting of critical path elements rather than comprehensive event chains, seeking visual emphasis on attack-relevant activities.

Academic research on provenance graph reduction for visualization remains scarce. Existing techniques prioritize storage optimization—preserving complete raw event histories for forensic completeness—rather than cognitive optimization for human comprehension. More aggressive reduction techniques that prune information flow edges providing minimal investigative value could distill visualizations to key steps representing underlying attack behavior. Such approaches would enable better integration into investigation workflows by surfacing actionable intelligence while suppressing extraneous technical detail, though they necessarily trade forensic completeness for analytical clarity.

2) Analysts Prefer Tabular Interfaces: Beyond visualization design challenges, practitioners consistently questioned their ability to query provenance data through familiar mechanisms. Security analysts develop investigation techniques through iterative querying of tabular datasets, and our graphical interface removed the tabular paradigm central to their training and daily practice. As one analyst stated, ”interacting with security logs in queries and a table is faster than trying to examine a graph.” This preference reflects established workflows where practitioners construct investigative queries against structured data—filtering, aggregating, and joining across log sources—rather than navigating visual representations through point-and-click interaction.

Most provenance-based academic research has presented graphical visualizations, implicitly assuming visual interfaces naturally convey causal relationships more effectively than tabular representations. However, encouraging adoption may require presenting causal information in formats that align with existing mental models and query-driven investigation patterns. This finding suggests that effective provenance tools must either provide tabular query interfaces over causal data structures—enabling SQL-like operations on dependency relationships—or fundamentally reconceptualize how provenance information integrates into query-driven workflows. Simply replacing familiar tabular paradigms with graph-based exploration may create adoption barriers regardless of the analytical advantages provenance provides, as the cognitive and procedural costs of interface transition outweigh perceived benefits

for time-constrained practitioners operating under pressure.

III. Why should the security industry be excited about data provenance?

Customer discovery reveals that while provenance requires additional research for commercial readiness, fundamental investigation and detection challenges stemming from tabular data architectures position provenance to deliver superior solutions that address core operational pain points. For example, current investigations rely on standardized playbooks to identify indicators of compromise surrounding alerts, a process constrained by predetermined query sequences. As one analyst explained, "identifying if a given alert is an attack within a given time frame is an art and may not always be right." SOCs accept this approach because playbooks standardize investigations and limit response times, enabling consistent analyst performance across experience levels, yet they sacrifice comprehensive visibility into threat actor behavior. Playbooks rarely reveal complete attack chains or expose the full scope of adversary activity within enterprise environments, enabling analysts to dismiss alerts that represent genuine threats as false positives based on incomplete information. Academic provenance research [11, 17] addresses this limitation directly through causal reconstruction of event sequences leading to alerts, automatically surfacing the complete attack picture by tracing information flow dependencies across system boundaries. Commercial adoption of such solutions would provide analysts with the information necessary for optimal decision-making, making enterprise's defense more robust.

Detection systems face parallel structural limitations imposed by tabular data architectures. Current rules operate two-dimensionally within tabular log formats, flagging isolated behaviors associated with malicious activity without understanding causal relationships between events. This constrained view generates excessive false positives that overwhelm analyst capacity, forcing SOCs to disable noisy rules entirely rather than tolerate alert fatigue. This compromise sacrifices detection coverage for operational efficiency, creating systematic blind spots to the very attacks those disabled rules were designed to identify. Provenance-based alert triage [12, 14] fundamentally resolves this dilemma by grouping causally related alerts into coherent attack narratives, enabling analysts to retain comprehensive detection coverage while dramatically reducing investigation burden through automated causal analysis. Advanced provenance intrusion detection systems [10, 15, 22] demonstrate that enriching detection with causal context achieves exceptional accuracy in identifying complete attack chains—whether known threat patterns or novel zero-day campaigns—fundamentally outperforming traditional correlation-based approaches that rely on heuristics.

There is tremendous potential value to be derived from provenance if SOCs were able to integrate it into their existing workflows. Commercial adoption, therefore, hinges on addressing key integration barriers that we identified through customer discovery, that enables a seamless deployment within existing investigation and detection pipelines. Having security analysts be able to quickly utilize provenance within to conduct their daily tasks with limited disruption is the key to provenance adoption.

IV. Discussion & Conclusion

Our customer discovery findings reveal critical opportunities for translational research that bridges the gap between "academic" provenance and commercial viability. Lesson II-A confirms that while academic motivational examples reflect genuine real-world phenomena, they manifest differently in practice. Organizations address alert fatigue by disabling detection rules rather than processing excessive alerts. Addressing this gap requires research accounting for current commercial practices, demonstrating provenance viability within reduced detection rule environments.

Lesson II-B identifies a knowledge translation barrier, highlighting industry communication complexity while suggesting valuable research directions on security practitioner mental models. We also document how industry-sourced solutions address operational problems through ad hoc approaches that diverge from classical information flow analysis – *but are they wrong for doing so?* While it is clear that commercial products have problems, direct comparisons to experimental provenance approaches are limited. Future research should pursue direct comparisons between provenance and commercial tooling in controlled experiments, considering "mundane" high volume attacks in addition to exciting and sophisticated APT's.

Lesson II-C exposes enterprise security stacks as fractured ecosystems with requirements diverging substantially from academic solution assumptions. Prior provenance-based auditing system design research warrants revisiting to incorporate this operational reality. Future research could develop robust provenance-based techniques maintaining effectiveness despite missing information, better reflecting current ecosystems.

Lesson II-D reveals the absence of research on threat investigation user interfaces for provenance analysis, representing an opportunity for the community to advance toward grounded understanding of security analyst experiences. Future research must understand how security practitioners utilize tools and how provenance-based analysis integrates into workflows, potentially requiring redesigned visualization approaches.

Deeper understanding of commercial software has the potential to inform and improve scientific research approaches. For example, awareness of ad-hoc false alert reduction mechanisms, like UEBA, provides specificity needed for direct comparison of causal and correlational analysis in evaluations. Process tree visualization popularity in commercial products offers another opportunity—as process trees constitute provenance graph subsets, ablation experiments can straightforwardly compare analytical power. Understanding actual AI cybersecurity solution capabilities enables academics to engage skeptically with AI hype while measuring genuine operational utility. We encourage researchers to engage industry more frequently through entrepreneurial activity or other participation, as it provides substantial inspiration for basic scientific research.

V. Acknowledgment

We thank the anonymous reviewers for their constructive feedback to improve this paper. This work was supported by NSF TI 2424261 and the Illinois Proof-of-Concept (IPOC) program. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or the sponsors.

References

[1] Abdulellah Alsaheel, Yuhong Nan, Shiqing Ma, Le Yu, Gregory Walkup, Z. Berkay Celik, Xiangyu Zhang, and Dongyan Xu. ATLAS: A sequence-based learning approach for attack investigation. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3005–3022. USENIX Association, August 2021.

[2] Steve Blank. *The four steps to the epiphany: successful strategies for products that win*. John Wiley & Sons, 2020.

[3] Carbon Black. Global incident response threat report. https://www.carbonblack.com/global-incident-response-threat-report/november-2018/, November 2018. Last accessed 04-20-2019.

[4] Z. Cheng, Q. Lv, J. Liang, Y. Wang, D. Sun, T. Pasquier, and X. Han. KAIROS: Practical Intrusion Detection and Investigation using Whole-system Provenance. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 9–9, Los Alamitos, CA, USA, may 2024. IEEE Computer Society.

[5] Crowdstrike. Why Dwell Time Continues to Plague Organizations. https://www.crowdstrike.com/blog/why-dwell-time-continues-to-plague-organizations/, 2019.

[6] Feng Dong, Liu Wang, Xu Nie, Fei Shao, Haoyu Wang, Ding Li, Xiapu Luo, and Xusheng Xiao. {DISTDET}: A {Cost-Effective} distributed cyber threat detection system. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 6575–6592, 2023.

[7] FireEye, Inc. How Many Alerts is Too Many to Handle? https://www2.fireeye.com/StopTheNoise-IDC-Numbers-Game-Special-Report.html, 2019.

[8] Peng Gao, Xusheng Xiao, Zhichun Li, Fengyuan Xu, Sanjeev R Kulkarni, and Prateek Mittal. {AIQL}: Enabling efficient attack investigation from system monitoring data. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, pages 113–126, 2018.

[9] Helen Gibson, Joe Faith, and Paul Vickers. A survey of two-dimensional graph layout techniques for information visualisation. *Information visualization*, 12(3-4):324–357, 2013.

[10] A. Goyal, G. Wang, and A. Bates. R-CAID: Embedding Root Cause Analysis within Provenance-based Intrusion Detection. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 257–257, Los Alamitos, CA, USA, may 2024. IEEE Computer Society.

[11] Wajih Ul Hassan, Adam Bates, and Daniel Marino. Tactical Provenance Analysis for Endpoint Detection and Response Systems. In *41st IEEE Symposium on Security and Privacy (SP)*, Oakland'20, May 2020.

[12] Wajih Ul Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. NoDoze: Combatting Threat Alert Fatigue with Automated Provenance Triage. In *26th ISOC Network and Distributed System Security Symposium*, NDSS'19, February 2019.

[13] IBM Security. Cost of a Data Breach Report 2023, July 2022.

[14] Muhammad Adil Inam, Jonathan Oliver, Raghav Batta, and Adam Bates. Carbon filter: Real-time alert triage using large scale clustering and fast search. In *28th International Symposium on Research in Attacks, Intrusions and Defenses*, RAID'25, October 2025.

[15] Baoxiang Jiang, Tristan Bilot, Nour El Madhoun, Khaldoun Al Agha, Anis Zouaoui, Shahrear Iqbal, Xueyuan Han, and Thomas Pasquier. ORTHRUS: Achieving High Quality of Attribution in Provenance-based Intrusion Detection Systems. In *34th USENIX Security Symposium*, Sec'25. USENIX Association, August 2025.

[16] Sadegh M. Milajerdi, Birhanu Eshete, Rigel Gjomemo, and Venkat N. Venkatakrishnan. Propatrol: Attack investigation via extracted high-level tasks. In Vinod Ganapathy, Trent Jaeger, and R.K. Shyamasundar, editors, *Information Systems Security*, pages 107–126, Cham, 2018. Springer International Publishing.

[17] S. Momeni Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. Venkatakrishnan. Holmes: Real-time apt detection through correlation of suspicious information flows. In *40th IEEE Symposium on Security and Privacy*, Oakland'19, Los Alamitos, CA, USA, may 2019. IEEE Computer Society.

[18] Sadegh M Milajerdi, Birhanu Eshete, Rigel Gjomemo, and VN Venkatakrishnan. Poirot: Aligning attack behavior with kernel audit records for cyber threat hunting. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 1795–1812, 2019.

[19] Yutao Tang, Ding Li, Zhichun Li, Mu Zhang, Kangkook Jee, Xusheng Xiao, Zhenyu Wu, Junghwan Rhee, Fengyuan Xu, and Qun Li. Nodemerge: Template based efficient data reduction for big-data causality analysis. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pages 1324–1337, New York, NY, USA, 2018. ACM.

[20] Raga'ad M Tarawaneh, Patric Keller, and Achim Ebert. A general introduction to graph visualization techniques. In *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011*, pages 151–164. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2012.

[21] Zhang Xu, Zhenyu Wu, Zhichun Li, Kangkook Jee,

Junghwan Rhee, Xusheng Xiao, Fengyuan Xu, Haining Wang, and Guofei Jiang. High fidelity data reduction for big data security dependency analyses. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 504–516, New York, NY, USA, 2016. ACM.

[22] Jun Zengy, Xiang Wang, Jiahao Liu, Yinfang Chen, Zhenkai Liang, Tat-Seng Chua, and Zheng Leong Chua. Shadewatcher: Recommendation-guided cyber threat analysis using system audit records. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 489–506, 2022.

[23] Bo Zhang, Yansong Gao, Changlong Yu, Boyu Kuang, Zhi Zhang, Hyoungshick Kim, and Anmin Fu. TAPAS: An Efficient Online APT Detection with Task-guided Process Provenance Graph Segmentation and Analysis. In *34th USENIX Security Symposium*, Sec'25. USENIX Association, August 2025.