My research advances trustworthy cyber defense at the intersection of system security, graph machine learning (graph ML), and LLM-based Agentic systems. I connect real-world security constraints to fortify theoretical defenses: *can we build systems that not only detect and withstand adaptive adversaries, but also explain and govern their detection decisions in ways security analysts and practitioners can trust?*

In system security, we capture system event logs that trace information and control-flow dependencies. System provenance captures information about system resources (processes, files, and sockets) and their informational dependencies (read, write, and execute events). By examining these system *provenance logs (or provenance logs)*, we can monitor the behavior of all processes on a system and create a heterogeneous provenance graph by associating causal dependencies between these system events. These dependencies are critical for accurately representing the system. Fine-grained provenance graphs are uniquely positioned to help detect and thwart stealthy attacks. The fundamental questions that I explore are:

- *Robustness:* How do adversaries mutate actions under problem space constraints to evade detectors?
- *Detection:* How to model benign behavior under resource constraints without losing stealthy attack signals?
- *Explainability:* How do black-box models and LLM Agents produce analyst-understandable rationales?
- *Privacy:* How to release sensitive system data under formal privacy guarantees while preserving utility?

These problems are hard because heterogeneous, high-variance provenance logs make it difficult to build stable, benign profiles; adversaries adapt, so "universal" attack assumptions and one-size-fits-all defenses are unrealistic; edge/IoT deployments face strict compute and privacy constraints; and explanations must tie to verifiable events that analysts can act on. However, these very challenges also create a unique opportunity: provenance graphs expose rich causal structure that, if modeled and governed correctly, can support detectors that are simultaneously robust to adaptive attackers and explanation-ready for human analysts.

I have designed end-to-end frameworks that (i) harden Machine Learning (ML)-based detectors against adaptive adversaries, (ii) developed federated provenance-based intrusion detection system (PIDS) framework for the resource constrained IoT domain, (iii) made decisions transparent and actionable for analysts, and (iv) provided privacy guarantee on system provenance data using differential privacy (DP) framework. My research has produced adversarial robustness analysis of PIDS (ProvNinja [13]), federated PIDS framework (ProvIoT [11]) for IoT domain, interpretable explanations for Graph Neural Network (GNN)-based IDS (ProvExplainer [12]) and real-estate recommendations (Z-REx [8]), and a differentially-private data sharing pipeline for system provenance data (ProvDP [15]). My current research focuses on: (i) leveraging LLM agents (ProvSEEK [9]) for automated system-security forensics with verifiable, auditable responses, and (ii) evaluating the adversarial robustness of bot detectors (BoCloak [6]).

My contributions have appeared in top security and ML venues (e.g., USENIX Security, ACNS, and KDD), which include an oral paper at KDD 2025. I have always released open-source artifacts [5, 4, 14] for my research endeavors to promote reproducibility, which have been recognized by the community, with multiple stars and forks. I contributed heavily to open-source projects such as the Deep Graph Library (DGL) [10, 7, 1, 2], expanding their heterogeneous GNN explainability modules. My internship contributions include a patented explainability framework for a GNN-based recommendation system at Zillow Group, Inc., and the design of a location-dependent decryption cryptosystem, GeoGuard [3], at Ciholas, Inc.

Going forward, my lab will pursue three intertwined questions: (i) *Robustness*: modeling adaptive, domain feasible mutations in the problem space to develop realistic attack frameworks for domains using graph data beyond system logs, such as blockchain and social networks; (ii) *Detection*: advancing IDS pipelines by fusing LLM reasoning complemented with GNN inference on graph-structured system activity for robust anomaly detection and interpretable investigations; and (iii) *Explainability & Governance*: turning black-box models and agentic pipelines into grounded, analyst-ready rationales with policy-aware guardrails.

Research Directions

From Evasions to Evidence: A robustness deployment stack. My work in system security establishes a robust framework that ties adversarial analysis directly to problem-space and resource-aware constraints. This research exposes State-of-the-art (SOTA) PIDSs' blind spot using adversarial analysis and then converts those insights into protective, usable mechanisms for defenders, echoing a question-driven, impact-oriented approach. *ProvNinja* [13] delivers the first systematic bridge from feature-space adversarial attacks to *problem-space* system actions, showing that realizable, budgeted manipulations can evade SOTA PIDS, therefore, motivating rigorous robustness and adversarial testing regimes. Complementing this, *ProvIoT* [11] demonstrates a federated edge–cloud architecture that supports on-device anomaly detection across heterogeneous IoT fleets, keeping data local while leveraging global aggregation and maintaining high detection performance under strict hardware budgets. Together, these systems realize an auditable pipeline, so defenders can measure, harden, and validate PIDS with confidence.

Impact. SOTA PIDS operate on high-variance, workflow-dependent provenance graphs, where small real-world behavioral deviations can nonlinearly perturb features and introduce background noise; as a result, claims of robustness that hold only in the feature space fail to capture what attackers can actually achieve in real-world systems. But, before I could study stealthy attacks from Advanced Persistent Threat (APT) actors and Fileless Malware writers, I had to capture high-quality attack logs. I constructed isolated, automated testbeds that execute APT and Fileless malware scenarios and record trustworthy traces, translating statistical attack intents into concrete action sequences. By using a federated design, I brought the superior PIDS capability to resource-constrained devices, preserving privacy and availability while meeting stringent device constraints. This work also motivated me to develop differential-privacy guardrails, *ProvDP* [15], that enable institutions to share provenance-derived insights without leaking sensitive identities or workflows, ensuring that robustness and governance advance together.

Explainability for Deep Models: From black boxes to grounded rationales. To convert PIDS outputs into analyst-ready evidence, I design explanation methods that align with the needs of two distinct audiences: PIDS developers who debug architectures and features, and *users* (e.g., security analysts) who must validate and act on decisions. Explanations, therefore, must speak in the *problem space*, system events (represented in graphs as nodes, edges, subgraphs) that humans recognize, rather than only in terms of weights or gradients. *ProvExplainer* [12] instantiates this for PIDS by training an interpretable surrogate over security-aware discriminant subgraph patterns and graph-structural features, yielding instance-level rationales that map directly to system behaviors and APT kill-chain stages or Tactics, Techniques, and Procedures (TTPs). Our evaluation showed ProvExplainer performing superiorly across all explanability metrics (e.g., fidelity^{+}/fidelity^{-}/precision/recall and a human-actionability distance).

Extending these explanation principles beyond security, *Z-REx* introduces a human-interpretable explainer for GNN-based Real Estate recommendation systems (RecSys). It combines targeted feature perturbation with structural perturbation to ground-truth-aware subgraph rationales that Zillow Group teams can inspect and use. Z-REx was evaluated on real-world real estate data to show it delivers higher explanation fidelity than SOTA GNN explainers while remaining actionable for stakeholders. Concretely, ProvExplainer shows how security-aware patterns and structural features produce instance-level rationales that track real TTPs and outperform prior explainers on intrusion detection tasks, while Z-REx demonstrates a scalable, interpretable RecSys explanation.

Impact. Explainability in security and recommendation is difficult because the objects to be explained are (i) *structurally* complex (heterogeneous, multi-hop graphs where small real-world edits nonlinearly perturb features), (ii) *audience-mismatched* (developer-oriented internals vs. operator-oriented evidence), and (iii) *metric-contested* (fidelity alone misses whether an explanation is correct, verifiable, or useful to humans). My contribution is to close these gaps by (1) grounding explanations in problem-space artifacts (subgraphs, relations, and events) that practitioners can verify; (2) coupling fidelity with precision/recall against curated

ground truth and a human-actionability metric that measures investigative distance; and (3) identifying the explanation relevance beyond the security domain to the recommendations domain. These insights guide a unified program: evidence-linked explanations that remain faithful to the model's behavior (e.g., PIDSs' detection and RecSyss' recommendation) and are actionable for users.

Agentic Provenance Forensics: Grounded agentic reasoning using RAG I aim to address the critical challenge of ensuring that AI agents deployed in cybersecurity environments operate safely, reliably, and in alignment with human values and security requirements. *ProvSEEK* [9] reframes digital forensics as a *grounded*, LLM-powered Agentic workflow that fuses cyber–threat intelligence with provenance databases. The system plans schema-aware SQL probes via retrieval-augmented generation (RAG) context augmentation, runs bounded chain-of-thought (CoT) reasoning, and *verifies* each intermediate claim by tying it to concrete system events present in system databases. To keep reasoning reliable and auditable, ProvSEEK utilizes *role-specific* agents (e.g., Digital Forensics and Incident Response (*DFIR*) analyst for correlation, system-security expert for provenance validation, and safety auditor for guardrails) that coordinate a toolkit spanning CTI retrieval, type-aware artifact lookups, and evidence correlation. As new evidence is discovered, the follow-up planner adjusts the investigation plan in a controlled manner, using targeted, type-safe queries and deduplicating results. These mechanisms prevent context explosion and maintain a faithful, traceable, and token-efficient narrative. In evaluations, ProvSEEK surpasses strong RAG baselines for intelligence extraction and outperforms SOTA PIDS on provenance datasets for threat identification.

Impact. This research focuses on developing comprehensive safety guardrails and alignment mechanisms to address AI hallucinations, overconfident responses, bias mitigation, and actionable decision-making. This research also focuses on developing novel confidence calibration methods to address the unique challenges of security triage, where overconfident AI responses can directly impact system safety. My contribution is a verification-first, role-specialized design that treats the LLM as a coordinator of provenance-aware tools rather than an oracle. ProvSEEK is, to my knowledge, the first LLM-powered *agentic* framework for automated, provenance-driven forensic analysis and threat intelligence extraction. Empirically, this closes the gap between retrieval quality and actionable detection improvements over SOTA PIDS while preserving auditability for incident response teams. These principles drive my roadmap for auditable agents that learn and expand plans prudently as evidence is discovered, and keep claims tethered to concrete system events.

Future Research Directions

Agentic Digital Forensics and Incident Response (DFIR). Agentic DFIR is challenging because investigations must be *verifiable*, *scalable*, and *adaptive*. Naïve prompting or time window slicing cannot tame billions of provenance events stored on provenance database; long, multi-hop chains require evolving plans; and LLMs can hallucinate unless every step is grounded in database-backed facts. Building on my current work, I will extend an agentic RAG pipeline that (i) retrieves only task-relevant context from a vector store, (ii) leverages GraphRAG to capture causal structure among Indicators of Compromise (IOCs) and TTPs along the cyber kill chain, and (iii) uses unsupervised filtering to prune SQL result sets into concise, high-signal evidence. The agent will explicitly represent uncertainty, refine queries when signals conflict, and emit *verifiable investigation thoughts*-SQL statements, retrieved artifacts, and decision traces—suitable for audit and replay. The system will integrate with PIDS triage and hunt loops, advancing both capability and governance by enabling agents that reason over provenance graphs with measurable fidelity, bounded autonomy, and enforceable guardrails for safe use in security-critical environments.

Realistic Adversarial Attacks on GNNs and LLMs. My goal is to harden graph and LLM learning pipelines by closing the loop between adversarial testing, robust training, and interpretable evaluation. Building on my work of realizable evasions for PIDS, I will generalize realistic attack generators and red-teaming tools beyond system logs to blockchain and large social graphs, stress-testing both link- and node-level tasks under practical domain constraints to ensure that models fail transparently and surface analyst-useful rationales.

For GNNs, I will design and evaluate both problem-space *evasion* attacks (minimal, semantics-preserving edge/node edits that flip link/node predictions under budget and observability constraints) and *poisoning* attacks (training-time manipulations that degrade downstream detection or ranking), with measurement that quantify stealth, cost, and collateral impact. For LLMs, I will build a red-teaming suite that composes *prompt-injection* and *jailbreak* patterns with tool access and retrieval steps, tracing how untrusted inputs (documents, APIs, users) steer plans, thoughts, tools, and claims; the suite will include automatic detectors, guardrails, and repair loops to harden agent policies and orchestrations.

Long-Term Vision. My research converges toward a *Trustworthy and Explainable AI* idea: governed Agentic forensics, and robust and interpretable graph learning that reason with verifiable evidence. By unifying adversarial robustness, explainability, privacy, and governance, my lab will deliver cyber defense that is *auditable by default*: detectors that withstand adaptive adversaries and agents whose actions can be traced, justified, and improved.

Funding Opportunities. I anticipate pursuing funding from multiple sources that align with my agenda on trustworthy AI for system security. Within NSF, I will target NSF Secure and Trustworthy Cyberspace (SaTC) for robust PIDS and privacy-preserving data sharing; and NSF Information Integration and Informatics (III) to support human-centered, explainable agentic forensics. I will also explore funding opportunities from the Department of Energy's Office of Cybersecurity, Energy Security, and Emergency Response (CESER) through their Cybersecurity Research, Development, and Demonstration (RD&D) program for energy infrastructure resilience and threat detection; and from the National Institutes of Health for medical data security and health informatics through their Data Management and Sharing (DMSP) policy initiatives.

I will also pursue early-career mechanisms, including the NSF CAREER Award and opportunities with DARPA, to develop adaptive red-teaming suites for GNNs and grounded, auditable LLM-guided forensics, and with the Office of Naval Research and the Air Force Office of Scientific Research to advance formal robustness and privacy guarantees for mission-critical cyber defense. In industry, I plan to apply to Google Research, Amazon, Microsoft, and Meta for projects on private graph learning, trustworthy explanations, and operationalizing security solutions at scale.

Leadership, Collaboration & Service

Leadership. During my PhD, I owned the end-to-end arc from data-use agreements and IRB coordination to pipeline design, evaluation, and publication. Across partnerships, problem definitions, evaluation protocols, and success metrics were co-designed with practitioners, yielding deployable insights rather than lab-only benchmarks. I have also mentored junior researchers in ML methodologies for System Security and in explanation methods, guiding research directions, experimental design, dataset selection, code reviews, and experimental audits, culminating in co-authored publications and artifact-ready releases.

Collaboration. These collaborations showed me that the same core problem can demand very different solutions under academic, industrial, and regulatory constraints, and taught me how to balance those perspectives to reach a shared solution that meets everyone's expectations. I have led multi-institution efforts across multiple countries (e.g., the United States, Singapore, and China), spanning academia (e.g., UT Dallas, Virginia Tech, UIUC, University of Central Oklahoma, ASTAR Institute for Infocomm Research, and Shanghai Jiao Tong University) and industry (e.g., Zillow Group, AWS, Acronis, NEC Labs, and Stellar Cyber). These collaborations translated to both academic papers and patent publications.

Service. I serve the community as a reviewer for security venues (e.g., USENIX Security, CCS, ACM Computing Surveys, ACM Transactions on Privacy and Security (TOPS), IEEE Transactions on Information Forensics and Security(TIFS), and IEEE Transactions on Dependable and Secure Computing (TDSC)) and machine-learning venues (ICML, ICLR, KDD, and Learning on Graphs). I have also served as an Artifact Evaluator for both security (e.g., USENIX Security, NDSS) and ML systems venues (e.g., MLSys, MobiSys).

References

[1] Kunal Mukherjee. Implemented pgexplainer for homogeneous graph. https://github.com/dmlc/dgl/pull/5550, 2023. Accessed: November 6, 2023.

[2] Kunal Mukherjee. Implemented subgraphx explainer for homogeneous graph. https://github.com/dmlc/dgl/pull/5315, 2023. Accessed: November 6, 2023.

[3] Kunal Mukherjee. Geoguard: Uwb timing-encoded key reconstruction for location-dependent, geographically bounded decryption. arXiv preprint / manuscript, 2025. Under submission; preprint available.

[4] Kunal Mukherjee. Proviot (github repository). https://github.com/syssec-utd/proviot, 2025. Accessed: October 22, 2025.

[5] Kunal Mukherjee. Provninja (github repository). https://github.com/syssec-utd/provninja, 2025. Accessed: October 22, 2025.

[6] Kunal Mukherjee, Zulfikar Alom, Tran Gia Bao Ngo, Cuneyt Gurcan Akcora, and Murat Kantarcioglu. Optimal transport-guided adversarial attacks on graph neural network-based bot detection. arXiv preprint / manuscript, 2026. Under submission; preprint available.

[7] Kunal Mukherjee and Nicholas Baker. Implemented pgexplainer for heterogeneous graph. https://github.com/dmlc/dgl/pull/5739, 2023. Accessed: November 6, 2023.

[8] Kunal Mukherjee, Zachary Harrison, and Saeid Balaneshin. Z-rex: Human-interpretable gnn explanations for real estate recommendations. In *KDD Workshop on Machine Learning on Graphs in the Era of Generative AI (MLoG-GenAI)*, Toronto, Canada, 2025. Oral presentation.

[9] Kunal Mukherjee and Murat Kantarcioglu. Llm-driven provenance forensics for threat intelligence and detection. arXiv preprint / manuscript, 2025. Under submission; preprint available.

[10] Kunal Mukherjee and Tianhao Wang. Heterogeneous graph support for gnnexplainer. https://github.com/dmlc/dgl/pull/4401, 2022. Accessed: November 6, 2023.

[11] Kunal Mukherjee, Joshua Wiedemeier, Qi Wang, Junpei Kamimura, John Junghwan Rhee, James Wei, Zhichun Li, Xiao Yu, Lu-An Tang, Jiaping Gui, and Kangkook Jee. Proviot: Detecting stealthy attacks in iot through federated edge-cloud security. In *Applied Cryptography and Network Security (ACNS)*, LNCS 14585, pages 241–268. Springer, 2024.

[12] Kunal Mukherjee, Joshua Wiedemeier, Tianhao Wang, Muhyun Kim, Feng Chen, Murat Kantarcioglu, and Kangkook Jee. Interpreting gnn-based ids detections using provenance graph structural features. 2023. Under submission; preprint available.

[13] Kunal Mukherjee, Joshua Wiedemeier, Tianhao Wang, James Wei, Feng Chen, Muhyun Kim, Murat Kantarcioglu, and Kangkook Jee. Evading provenance-based ml detectors with adversarial system actions. In *Proceedings of the 32nd USENIX Security Symposium*, Anaheim, CA, USA, 2023.

[14] Kunal Mukherjee and Jonathan Yu. prov-dp (github repository). https://github.com/provdp/prov-dp, 2025. Accessed: October 22, 2025.

[15] Kunal Mukherjee, Jonathan Yu, Partha De, and Dinil Mon Divakaran. Provdp: Differential privacy for system provenance dataset. In *Applied Cryptography and Network Security (ACNS)*, 2025.