CNS
CENTER FOR NETWORKED SYSTEMS

CNS Research Review Agenda, Thursday, May 8, 2025

8:30 to 9:00 a.m. | Name Badge Pick-Up and Breakfast
Computer Science and Engineering building at UC San Diego, Room 1242

9:00 a.m. | Welcome and Round Table Introductions [CSE Room 1242]
Stefan Savage and George Porter, CNS Co-Directors and CSE Professors

Discussion Group Introduction | Alisha Ukani, CNS/CSE Graduate Student

9:30 a.m. | Session Chair - Stefan Savage

- **Characterizing the MrDeepFakes Sexual Deepfake Marketplace (30 min.)**

Presenter: Deepak Kumar, CNS/CSE Faculty

- **Practical Support for Integrity Validation of Criminal Legal Process (15 min.)**

Presenter: Alisha Ukani, CNS/CSE Graduate Student (Alex Snoeren)

- **Count of Monte Crypto (15 min.)**

Presenter: Alex Liu, CNS/CSE Graduate Student (Stefan Savage and Geoff Voelker)

10:30 a.m. | Break and Informal Interaction (15 minutes)

10:45 a.m. | Session Chair - Pat Pannuto

- **Inferring Hospital Ransomware Outages (15 min.)**

Presenter: Sumanth Rao, CNS/CSE Graduate Student (Stefan Savage and Geoff Voelker)

- **Understanding the Efficacy of Phishing Training in Practice (30 min.)**

Presenter: Elisa Luo, CNS/CSE Graduate Student (Stefan Savage and Geoff Voelker)

- **CRASHCART: Resuscitating Hospitals Paralyzed by Ransomware (15 min.)**

Presenter: Kartikeyan Subramanyam, Network Systems Engineer, UC San Diego Health

11:45 a.m. | Group Photograph [Bear sculpture outside of CSE]

12:00 p.m. | Discussion Groups Meet (30 minutes) [CSE Rooms 4262, 3127, 2217]
Small discussion groups where students and industry members can discuss open research questions on various topics.

12:45 p.m. | Lunch [Jacobs Hall, Room 1108]

CNS
CENTER FOR NETWORKED SYSTEMS

CNS Research Review Agenda (Page 2), Thursday, May 8, 2025

2:00 p.m. | Session Chair - George Porter [CSE Room 1242]

- **Alarming Alarms: Analyzing the Security of Dealer-Installed Car Alarm Systems (30 min.)**

Presenter: Yibo Wei, CNS/CSE Graduate Student (Aaron Schulman)

- **Don't Look Up, There Are Sensitive Internal Links in the Clear on GEO Satellites (30 min.)**

Presenter: Morty Zhang, CNS/CSE Graduate Student (Aaron Schulman)

- **Type Enforced Peripheral Management (30 min.)**

Presenter: Tyler Potyandy, CNS/CSE Graduate Student (Pat Pannuto)

3:30 p.m. | Break and Informal Interaction (30 minutes)

4:00 p.m. | Session Chair - Alex C. Snoeren

- **Automated Verification 101: How Verification Impacts Software Engineering (30 min.)**

Presenter: Eric Mugnier, CNS/CSE Graduate Student (YY Zhou)

- **Mustin: Taming Multi-Tier-SLO LLM Serving with Adaptive Parallelism (30 min.)**

Presenter: Vikranth Srivatsa, CNS/CSE Graduate Student (Yiying Zhang)

5:00 p.m. | Lightning Talks – Graduate Students and Industry Representatives (30 min.)

6:00 p.m. | Graduate Student Poster Session and Dinner Reception [Great Hall @ UCSD]

-See Day Two on the Next Page-

CNS
CENTER FOR NETWORKED SYSTEMS

Day Two / CNS Research Review Agenda, Friday, May 9, 2025

9:00 to 9:30 a.m. | Name Badge Pick-Up and Breakfast
[Computer Science and Engineering building at UC San Diego, Room 1242]

9:30 a.m. Session Chair | Earlence Fernandes [CSE Room 1242]
- **Skylink: Assessing the Viability of Internet-of-Things Coverage from Aerial Vehicles (15 min.)**

Presenter: Alex Yen, CNS/CSE Graduate Student (Pat Pannuto)
- **Breaking Enterprise Network Authentication (30 min.)**

Presenter: Miro Haller, CNS/CSE Graduate Student (Nadia Heninger)
- **Wide-area Carbon Load of Network Transfers (15 min.)**

Presenter: Amanda Tomlinson, CNS/CSE Graduate Student (George Porter)

10:30 a.m. | Break and Informal Interaction (15 minutes)

10:45 a.m. | Session Chair - Geoff Voelker [CSE Room 1242]
- **Sublet Your Subnet: Inferring IP Leasing in the Wild (15 min.)**

Presenter: Ben Du, CNS/CSE Graduate Student (Alex C. Snoeren, kc claffy)
- **Fun-tuning: Characterizing the Vulnerability of Proprietary LLMs to Optimization-based Prompt Injection Attacks via the Fine-Tuning Interface (15 min.)**

Presenter: Andrey Labunets, CNS/CSE Graduate Student (Earlence Fernandes)
- **Deep-Learning-Driven Prefetching for Far Memory (15 min.)**

Presenter: Yutong Huang, CNS/CSE Graduate Student (Yiying Zhang)
- **Cognify: Supercharging Gen-AI Workflows with Hierarchical Autotuning (15 min.)**

Presenter: Zijian He, CNS/CSE Graduate Student (Yiying Zhang)

11:45 a.m. | Open Floor [Feedback from company representatives]

12:00 p.m. | Lunch and CNS Research Review Conclusion [CSE Room 1202]

-CNS Research Review Conclusion-

-See Abstracts on Next Page-

Abstracts in order of appearance:

Characterizing the MrDeepFakes Sexual Deepfake Marketplace - Deepak Kumar

The prevalence of sexual deepfake material has exploded over the past several years. Attackers create and utilize deepfakes for many reasons: to seek sexual gratification, to harass and humiliate targets, or to exert power over an intimate partner. In part enabling this growth, several markets have emerged to support the buying and selling of sexual deepfake material. In this paper, we systematically characterize the most prominent and mainstream marketplace, MrDeepFakes. We analyze the marketplace economics, the targets of created media, and user discussions of how to create deepfakes, which we use to understand the current state-of-the-art in deepfake creation. Our work uncovers little enforcement of posted rules (e.g., limiting targeting to well-established celebrities), previously undocumented attacker motivations, and unexplored attacker tactics for acquiring resources to create sexual deepfakes.

Practical Support for Integrity Validation of Criminal Legal Process - Alisha Ukani

In the US, criminal legal process (e.g., search warrants, subpoenas, etc.) is the key mechanism by which law enforcement entities and courts compel third parties (e.g., such as Google or Meta) to produce evidence or take actions in support of a criminal investigation. However, such orders are simply documents and there is no inherent mechanism to guarantee their integrity. In recent years there have been increasing reports of forged legal process used to advance a variety of criminal purposes, with online sellers now offering such capabilities for a fee.

Our work focuses on designing a practical system to help third parties identify when the legal orders they receive are the result of forgery or tampering. Based on a series of conversations with judges, court clerks, law enforcement and the legal compliance components of large online service providers, we have identified a set of security and functionality goals that capture both the integrity needs of this ecosystem and the reality of how legal process is created, modified and served. Based on these, we have designed a system, to be deployed at both issuers (e.g., courthouses) and providers (e.g., Google) to help authenticate and validate the integrity of such documents after they are served.

Count of Monte Crypto - Alex Liu

Between 2021 and 2023, crypto assets valued at over $US2.6 billion were stolen via attacks on "bridges"—decentralized services designed to allow inter-blockchain exchange. While the individual exploits in each attack vary, a single design flaw underlies them all: the lack of end-to-end value accounting in cross-chain transactions. In this paper, we empirically analyze over 10M million transactions used by key bridges during this period. We show that a simple invariant that balances cross-chain inflows and outflows is compatible with legitimate use, yet precisely identifies every known attack (and several likely attacks) in this data. Further, we show that this approach is not only sufficient for post-hoc audits, but can be implemented in-line in existing bridge designs to provide generic protection against a broad array of bridge vulnerabilities.

Inferring Hospital Ransomware Outages - Sumanth Rao

Ransomware attacks affecting hospitals pose a substantial risk to patient safety, as well as the financial and operational viability of the national healthcare infrastructure. Current regulatory framework and status quo behavior result in delayed reporting of ransomware attacks by the hospitals themselves. This causes a detrimental effect on the ability to rapidly deploy resources to affected institutions, and in preparing unaffected hospitals within the "cyber blast radius" for the increased patient diversion and workflow burdens. This talk presents an ongoing effort to develop a measurement tool which can be used to infer ransomware-induced hospital network outages. Using this, we characterize ransomware attacks in the past and other outages affecting hospital organizations.

Understanding the Efficacy of Phishing Training in Practice - Elisa Luo

Our work empirically evaluates the efficacy of two ubiquitous forms of enterprise security training: annual cybersecurity awareness training and embedded anti-phishing training exercises. Specifically, our work analyzes the results of an 8-month randomized controlled experiment involving ten simulated phishing campaigns sent to over 19,500 employees at a large healthcare organization. Our results suggest that these efforts offer limited value. First, we find no significant relationship between whether users have recently completed cybersecurity awareness training and their likelihood of failing a phishing simulation. Second, when evaluating recipients of embedded phishing training, we find that the absolute difference in failure rates between trained and untrained users is extremely low across a variety of training content. Third, we observe that most users spend minimal time

interacting with embedded phishing training material in-the-wild; and that for specific types of training content, users who receive and complete more instances of the training can have an increased likelihood of failing subsequent phishing simulations. Taken together, our results suggest that anti-phishing training programs, in their current and commonly deployed forms, are unlikely to offer significant practical value in reducing phishing risks.

CRASHCART: Resuscitating Hospitals Paralyzed by Ransomware - Kartikeyan Subramanyam
Ransomware attacks on hospitals are not white collar crimes, they are threat-to-life crimes because they directly threaten a hospital's ability to provide patient care, which puts patient safety at risk. CRASHCART is a modularly developed, rapidly deployable, backup Internet access and clinical functionality solution with a plethora of tools to deploy in hostile environments, with a particular use case centered on hospitals hit by ransomware. Our primary goal was to build a mobile, rapidly deployable and robust network, as well as associated infrastructure to facilitate the connected clinical workflows imperative for critical patient care.

Alarming Alarms: Analyzing the Security of Dealer-Installed Car Alarm Systems - Yibo Wei
Our analysis of three popular aftermarket alarm ecosystems revealed a critical vulnerability in one major system that allows remote attackers to control vehicle functions including the immobilizer, doors, and CAN bus. With an estimated 1.68 million vulnerable alarms deployed, we recommend combining security design elements from multiple systems to achieve defense in depth.

Don't Look Up, There Are Sensitive Internal Links in the Clear on GEO Satellites - Morty Zhang
Geosynchronous (GEO) satellite links provide IP backhaul to remote critical infrastructure for utilities, telecom, government, military, and commercial users. GEO users benefit from reliable coverage, which blankets thousands of kilometers. Unfortunately, this also makes GEO links a desirable target for passive interception attacks. An attacker can passively scan the sky from a single location and observe thousands of backhaul links across a continent. Given their sensitivity and ease of interception, one would expect the traffic on GEO data links to be encrypted. In this work, we show there are many unencrypted satellite backhaul links that are carrying private network traffic for critical use cases. This is exposing sensitive data on internal network links. We found these links by performing the first broad scan of IP traffic on 39 GEO satellites across 25 distinct longitudes with 411 transponders. This required a new GEO satellite traffic parser that can handle many different network protocol stacks used by heterogeneous endpoint equipment. We found that 50% of GEO links lacked link-layer encryption and contained cleartext IP traffic, even though encryption has been standard practice for GEO satellite television for decades. Further, we found this internal traffic is also not encrypted at the network layer or above, offering a unique perspective into the weak security practices of organizations on links assumed to be internal. Specifically, we observed cleartext traffic for cellular backhaul from remote cell sites of several providers (i.e., calls, texts), job scheduling for an electrical utility, military asset tracking, and inventory management for global retail stores.

Type Enforced Peripheral Management - Tyler Potyondy
Low-level software drivers underpin the digital world and are used in a range of safety and security critical systems--necessitating their correctness. To ensure the correctness of these systems, driver developers must guarantee that all hardware-software interactions conform to the hardware's specification. We introduce a principled approach to modeling hardware using a formalization of hardware states. Our system utilizes type-state programming and the Rust compiler in conjunction with our hardware state formalization to statically determine the correctness of hardware-software interactions. Practically, this enables developers to more easily represent hardware specifications. We evaluate our system across two hardware manufacturers and multiple drivers, statically enforcing hardware-software interaction correctness while imposing minimal overheads in execution time and code size.

Automated Verification 101: How Verification Impacts Software Engineering - Eric Mugnier
Verification has never been as popular as today. While the excitement is palpable, feedback on the practical use of verifiers remains scarce. We lack a clear understanding of how these tools are employed, the challenges users encounter, and which types of projects are well-suited for verification. Although some insights can be gleaned from blog posts, and talks, these are individual anecdotes rather than a comprehensive view of the field.
In our work, we address these gaps by interviewing several industrial users of verification tools, focusing on how the use of an automated verifier–specifically Dafny– impacted the software development process. Using a method inspired by the sociology technique known as grounded theory, which develops theories and findings

directly from data, we aim to help the community better understand and effectively utilize these tools. Additionally, we interviewed the developers of Dafny to gain their perspective on the tool's use, lessons learned, and potential future directions. Through this work, we aim to provide a comprehensive view of automated verifiers, including insights on suitable projects, best practices and pitfalls, and motivations for future research.

Mustin: Taming Multi-Tier-SLO LLM Serving with Adaptive Parallelism - Vikranth Srivatsa

Modern LLM applications demand diverse service-level objectives (SLOs), ranging from real-time queries that require sub-200ms responses, to mid-tier tasks with second- to minute-level generation, and background jobs with best-effort, relaxed deadlines. Existing LLM serving systems either ignores SLO requirements or target a single SLO tier by applying request scheduling and batching techniques on a static and uniform GPU cluster configuration. As such, they cannot adapt to tiers of SLOs and workload dynamism.

In this talk, I will introduce our latest work, Mustin, a distributed LLM serving system that introduces dynamic tensor parallelism (TP) for tiered SLO management. Mustin dynamically configures the GPU cluster with TP levels based on the request SLOs and performance characteristics. To maximize SLO attainment and GPU efficiency, Mustin incorporates a set of novel mechanisms for efficienct TP level switching, including new computation and state migration techniques. We also introduce a new SLO-oriented scheduling policy and an efficient implementation of it involving a global scheduler and a set of local schedulers.

Skylink: Assessing the Viability of Internet-of-Things Coverage from Aerial Vehicles - Alex Yen

We provide one of the largest, terrestrial measurement studies on the LoRa wireless technology via aerial vehicles (i.e. weather balloons, airplanes). By measuring a LoRa Wide Area Network—the Helium Network—from the sky, we present results that indicate a viable means to achieve pervasive, Internet of Things (IoT) coverage today. Our work aims to extrapolate—at scale, the effectiveness of aerial, infrastructure networking for IoT, wireless coverage

Breaking Enterprise Network Authentication - Miro Haller

The RADIUS protocol is the de facto standard lightweight protocol for authentication, authorization, and accounting for networked devices. It is used to support remote access for diverse use cases including network routers, industrial control systems, VPNs, enterprise Wi-Fi including the Eduroam network, Linux Pluggable Authentication Modules, and mobile roaming and Wi-Fi offload. This talk presents the Blast-RADIUS vulnerability which allows a machine-in-the-middle attacker on RADIUS to authenticate themselves to a device as an arbitrary user with privileges of their choosing. We discuss the proof of concept applications of our attack against popular RADIUS implementations, and the large-scale disclosure process and mitigation efforts in collaboration with CERT and IETF.

Wide-Area Carbon Load of Network Transfers - Amanda Tomlinson

Demand continues to grow for Internet resources, and this demand is driving a rise in energy consumption and carbon emissions. A large amount of research has gone into reducing datacenter emissions, with relatively little work on the wide area network. In this talk we look into the dynamics of network carbon emissions by joining public traceroute datasets with carbon emissions data.

Sublet Your Subnet: Inferring IP Leasing in the Wild - Ben Du

In this talk, we analyze the IPv4 leasing ecosystem by designing a methodology to infer leased address space globally and study its impact on routing and hosting security. We infer that 4.1% of all advertised IPv4 prefixes (0.9% of routed v4 address space) were leased in April 2024. We show that leased address space is five times more likely to be abused compared to non-leased space.

Fun-tuning: Characterizing the Vulnerability of Proprietary LLMs to Optimization-based Prompt Injection Attacks via the Fine-Tuning Interface - Andrey Labunets

We surface a new threat to closed-weight Large Language Models (LLMs) that enables an attacker to compute optimization-based prompt injections. Specifically, we characterize how an attacker can leverage the loss-like information returned from the remote fine-tuning interface to guide the search for adversarial prompts. The fine-tuning interface is hosted by an LLM vendor and allows developers to fine-tune LLMs for their tasks, thus providing utility, but also exposes enough information for an attacker to compute adversarial prompts. Through an experimental analysis, we characterize the loss-like values returned by the Gemini fine-tuning API and

demonstrate that they provide a useful signal for discrete optimization of adversarial prompts using a greedy search algorithm. Using the PurpleLlama prompt injection benchmark, we demonstrate attack success rates between 65% and 82% on Google's Gemini family of LLMs. These attacks exploit the classic utility-security tradeoff - the fine-tuning interface provides a useful feature for developers but also exposes the LLMs to powerful attacks.

Deep-Learning-Driven Prefetching for Far Memory - Yutong Huang

Modern software systems face increasing runtime performance demands, particularly in emerging architectures like far memory, where local-memory misses incur significant latency. While machine learning (ML) has proven effective in offline systems optimization, its application to high-frequency, runtime-level problems remains limited due to strict performance, generalization, and integration constraints. We present DeepPF, a Linux-based far-memory system that leverages deep learning (DL) to efficiently perform accurate data prefetching. DeepPF separates application semantics from runtime memory layout, allowing offline-trained DL models to predict access patterns using a compact vocabulary of ordinal possibilities, resolved at runtime through lightweight mapping structures. By combining asynchronous inference, lookahead prediction, and a cache-resident DL model, DeepPF achieves high prediction accuracy with low runtime overhead. Our evaluation of DeepPF on four data-intensive workloads shows that it outperforms the state-of-the-art far-memory system by up to 3.6 times. Overall, this work demonstrates the feasibility and advantages of applying modern ML techniques to complex, performance-critical software runtime problems.

Cognify: Supercharging Gen-AI Workflows With Hierarchical Autotuning - Zijian He

Today's gen-AI workflows that involve multiple ML model calls, tool/API calls, data retrieval, or generic code execution are often tuned manually in an ad-hoc way that is both time-consuming and error-prone. In this paper, we propose a systematic approach for automatically tuning gen-AI workflows. Our key insight is that gen-AI workflows can benefit from structure, operator, and prompt changes, but unique properties of gen-AI workflows require new optimization techniques. We propose AdaSeek, an adaptive hierarchical search algorithm for autotuning gen-AI workflows. AdaSeek organizes workflow tuning methods into different layers based on the user-specified total search budget and distributes the budget across different layers based on the complexity of each layer. During its hierarchical search, AdaSeek redistributes the search budget from less useful to more promising tuning configurations based on workflow-level evaluation results. We implement AdaSeek in a workflow autotuning framework called Cognify and evaluate Cognify using six types of workflows such as RAG-based QA and text-to-SQL transformation. Overall, Cognify improves these workflows' generation quality by up to 2.8x, reduces execution monetary cost by up to 10x, and reduces end-to-end latency by 2.7x.