Research Statement

Arpit Gupta, Assistant Professor, UCSB

Summary of Major Research Accomplishments

My research has centered on **democratizing access to critical data** for both data-driven policymaking and trustworthy machine learning (ML). By building tools that provide access to the "right" data at scale, I've empowered policymakers to make more informed policy decisions and enabled ML-powered network automation for small and medium-scale enterprises.

The **Broadband-Plan Querying Tool (BQT)** is a prime example of this effort. BQT allows policymakers to access granular broadband data. This tool not only contributed to enhancing existing, often noisy broadband quality datasets [10] but also synthesizing the first comprehensive broadband affordability dataset [9]. The graduate student leading this effort received the prestigious **SIGCOMM Doctoral Dissertation Award** in 2024—highest recognition for a doctoral dissertation in communications and networking. More recently, we demonstrated how to use BQT to enable post hoc evaluation of multi-billion policy interventions. Specifically, our recent work audits the self-reported information provided by ISPs regarding the $10B Connect America Fund (CAF) program [8]. This study was also published as a Harvard Law Review article [7] and was the basis of our amici curiae brief at the US Supreme Court [4]. We are currently in the process of litigating against the concerned ISPs under the False Claims Act. Collectively, these endeavors have helped bridge fundamental data gaps to bring more transparency and diligence to multi-billion-dollar policy programs.

My research also made contributions in addressing the fundamental data gaps that affect the trustworthiness of ML models in networking. Specifically, I have realized a closed-loop ML workflow that enables iterative curation of the "right" data for a given learning problem and target environment. This effort has resulted in three open-sourced systems: **Pinot** [2], **netUnicorn** [1, 5], and **Trustee** [3, 6][1]. I am currently also working as a **Faculty Scientist at Berkeley Lab** to develop production-ready ML models for network operations.

Ongoing Efforts

My ongoing research explorations aim to answer the following research questions:

How to fundamentally advance broadband quality assessments using AI/ML? Despite improvements in broadband availability (e.g., National Broadband Map) and affordability data (e.g., BQT), there is still a critical gap in high-fidelity broadband quality assessments at scale. Current methods rely on speed tiers and aggregate QoS metrics (e.g., speed, latency, jitter) from crowdsourced tools, which fail to capture the actual user experience, especially under dynamic network conditions like congestion. As a result, policymakers are left with sparse and unrepresentative data to make informed decisions.

To address this, my research focuses on leveraging AI/ML to fundamentally improve broadband quality assessments. I am developing a network measurement framework that scales data collection and provides more accurate, context-aware metrics. Central to this approach is the creation of network foundation models, **netFound**, designed to learn dynamic representations of network data by exploiting large volumes of unlabeled data. These models will generate dynamic network representation (DNR) vectors from active measurements, effectively capturing how protocols perform under varying network conditions.

Once validated, these DNR vectors will be used to synthesize Quality of Experience (QoE) metrics for applications like YouTube and Zoom, simulating user experiences under similar network conditions. This AI-driven approach aims to bridge the gap between measured and experienced network quality, offering policymakers more reliable, real-world data to assess broadband performance—especially after large-scale infrastructure investments.

How to enable secure and performant community networks using AI/ML? Despite significant investments by the U.S. government to achieve *universal access*, many communities remain underserved or unserved. A key reason for this persistent gap is the centralized nature of these initiatives. Much of the funding is directed to large incumbent ISPs, whose priorities do not always align with the government's goal of reducing access inequity. Community networks—decentralized, community-driven infrastructures—have emerged as an alternative to the control exerted by these large ISPs. Currently, over 900 community networks

[1] This work also received the prestigious IETF's Applied Networking Research Prize

serve marginalized and underrepresented populations across the U.S., with the BEAD program aiming to expand these efforts. However, through discussions with community network operators, I have recognized a critical barrier: limited resources and a lack of technical expertise in rural and remote areas prevent these networks from providing services on par with larger ISPs.

To address this challenge, I propose exploring how advances in AI/ML and communication technologies can enable the design, prototyping, and deployment of resilient, self-managing community networks. These networks would allow non-expert operators to deploy, maintain, and troubleshoot local infrastructure autonomously. My work will investigate whether AI-powered systems can help bridge the technical expertise gap, making community networks both sustainable and capable of offering competitive services. This approach mirrors my current role at Berkeley Lab, where I am tasked with realizing AI-powered network operations for ESnet. I plan to use this experience to enable AI-powered community networks that present a more extreme case, with far greater constraints in terms of available expertise and resources.

If successful, this research could fundamentally decentralize the management of last-mile networks in the US and beyond, empowering smaller communities to take control of their broadband offerings and driving socioeconomic growth. It would also enable underprivileged communities to better integrate into the AI-driven digital economies of the future.

Planned Research Program

Looking forward, I aim to leverage my expertise to design tools and infrastructures that address the growing challenges of digital inequity in the age of AI and machine learning (ML). My research will aim to answer the following three key questions:

How to democratize access to AI and ML? Building on my prior work on broadband access, I aim to address the emerging inequities in access to AI/ML services. As AI/ML becomes increasingly pervasive, unequal access to these technologies—whether in precision medicine or LLMs—could mirror the inequalities we currently see in broadband access. My research will examine the intersection of broadband inequities and access to AI/ML, focusing on the networking requirements of AI/ML applications and identifying regions and demographic groups that are most vulnerable to exclusion. A key part of this work will analyze whether these underserved populations overlap with those identified by the FCC as lacking broadband access and whether current definitions of "underserved" need to be updated to reflect the demands of AI/ML-driven digital infrastructure. Additionally, I will investigate the redesign of last-mile and middle-mile infrastructure to meet the growing demand for AI/ML services, emphasizing the role policymakers must play in ensuring equitable access for all communities.

How to enable AI/ML for social good? The rapid development of large language models (LLMs) has propelled machine learning (ML), especially natural language processing (NLP), into the public spotlight. However, the transformative potential of AI extends well beyond NLP, with significant implications in fields such as precision medicine and governance. Centralizing AI/ML technologies in critical areas raises concerns about unequal access, which could exacerbate societal disparities. Efforts like the Trillion Parameter Consortium (TPC) are working to address the need for domain-specific, large-scale foundation models to drive scientific discovery.

My research will focus on overcoming key challenges in developing these models for domains such as education, governance, precision medicine, and climate science. By leveraging my expertise in developing AI/ML tools for networking—a data-starved application domain that entails high-stakes decision-making—and collaborating with domain experts, I aim to identify common workflows and build tools and infrastructures that lower the barriers to creating foundation models for new application domains, ensuring equitable access to these transformative technologies. Furthermore, I will explore the role of new data sources (e.g., wearable sensors) and advanced architectures (e.g., attention-based models) in facilitating these applications while addressing the socio-technical challenges of democratizing AI. My ultimate goal is to foster a more inclusive AI landscape that benefits underserved communities and advances social good.

References

[1] netunicorn: A network data-collection platform—https://netunicorn.github.io/. (Cited on page 1.)

[2] Pinot: A programmable infrastructure for ai/ml for networking—https://pinot.cs.ucsb.edu/. (Cited on page 1.)

[3] Trustee: A model explainability framework for Networking—https://trusteeml.github.io/. (Cited on page 1.)

[4] E. Belding, A. Gupta, and T. N. Narechania. Brief of Amici Curiae in Support of Respondents, Wisconsin Bell v. US ex rel. Heath (No. 23-1127)(US Oct. 1, 2024). *US ex rel. Heath*, (23-1127), 2024. (Cited on page 1.)

[5] R. Beltiukov, W. Guo, A. Gupta, and W. Willinger. In search of netunicorn: A data-collection platform to develop generalizable ml models for network security problems. In *CCS*, 2023. (Cited on page 1.)

[6] A. S. Jacobs, R. Beltiukov, W. Willinger, R. A. Ferreira, A. Gupta, and L. Z. Granville. Ai/ml for network security: The emperor has no clothes. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022.
IETF/IRTF Applied Networking Research Prize
Best Paper Honorable Mention. (Cited on page 1.)

[7] H. Manda, V. Srinivasavaradhan, L. Koduru, K. Zhang, X. Zhou, U. Paul, E. Belding, A. Gupta, and T. Narechania. Measuring broadband policy success, harvard law review. (Cited on page 1.)

[8] H. Manda, V. Srinivasavaradhan, L. Koduru, K. Zhang, X. Zhou, U. Paul, E. Belding, A. Gupta, and T. Narechania. Assessing the Efficacy of the Connect America Fund in Addressing Internet Access Inequities in the US. In *ACM SIGCOMM*, 2024. (Cited on page 1.)

[9] U. Paul, V. Gunasekaran, J. Liu, T. N. Narechania, A. Gupta, and E. Belding. Decoding the Divide: Analyzing Disparities in Broadband Plans Offered by Major US ISPs. In *ACM SIGCOMM*, 2023. (Cited on page 1.)

[10] U. Paul, J. Liu, M. Gu, A. Gupta, and E. Belding. The importance of contextualization of crowdsourced active speed test measurements. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC)*, 2022.
Distinguished Paper Award (Long). (Cited on page 1.)