Research Statement

Arpit Gupta, Assistant Professor, UCSB

In our modern world, the ability to access information and communication technology is not just a convenience but is increasingly seen as a crucial "human right". This ability is vital for people around the world to derive socio-economical benefits from the Internet by engaging in activities such as education, healthcare, commerce, and civic participation. However, despite years of effort from various stakeholders, a significant "digital divide" persists that sharply separates those with seamless access to the Internet and cutting-edge communication technologies from those who remain underserved. The wide-ranging social and economic consequences of this divide cannot be overstated.

Having experienced aspects of this digital divide on a personal level during my upbringing in an economically underdeveloped central India, I am aware of the societal consequences of this digital divide. I understand the importance of changing the status quo, and I am committed to pursuing a research agenda aimed at forging a path toward a more equitable digital future. In particular, as I embarked on my tenure-track position at UCSB, I reflected on how to best utilize my skills and training toward achieving this vision and identified two key areas where I felt I could make significant contributions toward realizing my vision.

The first area concerns Internet measurement research, with an emphasis on enabling **data-driven policymaking**. This approach centers around providing policymakers with access to the "right" data, which aids in evaluating existing policies and informing the syntheses of new policies. This effort is crucial for optimizing the use of limited capital resources to benefit underprivileged communities, thereby addressing their specific needs more effectively. The second area involves advancing Artificial Intelligence (AI) and Machine Learning (ML) for networking, with an emphasis on developing **production-ready AI/ML artifacts**. This effort is first and foremost concerned with collecting the "right" data for training ML models and creating related ML-based artifacts that facilitate the development of *self-driving* networks capable of safely operating production networks with minimal human intervention. Such artifacts would prove especially beneficial in network environments with limited budgets, operational capacity, and technical expertise, such as community networks serving underprivileged and underrepresented communities.

In this research statement, I will first discuss my contributions to enabling data-driven policymaking (Section 1) and developing production-ready AI/ML artifacts for networking (Section 2). Next, I will describe my ongoing research efforts and outline my long-term plans for contributing to an equitable digital future (Section 3).

1 Enabling Data-Driven (Broadband) Policymaking

Various stakeholders—including policymakers and grassroots-level non-profits—are committed to addressing Internet access and digital equity challenges to inform high-stakes policy interventions such as subsidy programs (e.g., the Affordable Connectivity Program, ACP), rate regulations (e.g., the Connect America Fund, CAF Program), and infrastructure funding (e.g., the Broadband Equity, Access, and Deployment, BEAD program). Among these, the BEAD program is poised to invest approximately $44 billion in enhancing broadband infrastructure for underserved areas within the United States. These programs are in critical need of high-quality data pertaining to broadband availability, quality, affordability, and adoption to inform funding decisions. For instance, the BEAD program requires the identification of regions where broadband is either (1) not available, (2) of substandard quality, or (3) unaffordable.

My foray into Internet measurement research initially involved using existing datasets to address a variety of policy questions. However, I soon realized that a significant gap exists between the data policymakers require and the type of data that is currently available to them. Often, the dataset at policymakers' disposal is sparse, noisy, or heavily reliant on questionable self-reported information from ISPs. As a result, reliance on these datasets can lead to misleading conclusions and flawed funding decisions, with underserved areas suffering the consequences. For example, the BEAD program's initial funding allocations, based on the FCC's National Broadband Map, a dataset criticized for its noisy and self-reported nature, have sparked serious concerns about the validity of the decisions made.

To address this disconnect and improve the reliability of broadband data for policy-making, my team developed the **Broadband-Plan Querying Tool (BQT)** [10]. It emulates a real user's interactions with ISP websites to gather data on advertised broadband speeds and rates at the level of street addresses. This tool represented a major breakthrough, democratizing access to data capturing how different ISPs offer broadband plans to their customers at the finest possible spatial granularity. Access to such a dataset

enabled answering policy questions that were unanswerable before. Specifically, it enabled the augmentation of widely available crowdsourced broadband quality data (e.g., Ookla's speed test data), demonstrating the importance of context in speed test measurements (IMC'22) [24]. It was also instrumental in creating the first comprehensive dataset on broadband affordability, shedding light on how broadband ISPs operate in quasi-competitive markets (SIGCOMM'23) [23]. These efforts have significantly empowered a diverse set of stakeholders, boosting their involvement in the $44B BEAD program's challenge process.

Augmenting existing crowdsourced speed test measurements (ACM SIGCOMM IMC'22 [24]). Platforms like Ookla's Speedtest and Measurement Lab's network diagnostic tool (NDT) are widely used to measure broadband speed from the user's perspective. While the data from these platforms is invaluable, interpreting it without appropriate context can lead to erroneous conclusions about broadband quality. The key challenge is differentiating whether a speed test result aligns with the user's subscribed plan. For instance, a speed test measurement of 10 Mbps download speed is reasonable if the user subscribed to a lower-tier 10 Mbps plan but is problematic if the user is subscribed to a high-speed 500 Mbps plan.

We developed a methodology that infers the subscription tier from crowdsourced measurements, applying this to characterize biases in public (FCC's MBA, Measurement Lab) and private (Ookla) datasets. Our empirical analysis highlights how a lack of context leads to misconceptions, and we provided recommendations for vendors and the FCC to incorporate contextual data for accurate interpretation. Our contributions were recognized with the **ACM SIGCOMM IMC Distinguished (Long) Paper Award**. In the immediate follow-up of this work, we are collaborating with team at Google's Measurement Lab (Christophe Diot and Phillipa Gill) to explore the possibility of inferring additional measurement contexts. Our aim is to identify additional contextual information, such as the nature of the last-hop links (i.e., wireless vs. wired) and the types of devices being used (i.e., laptops vs. mobile devices). This effort is intended to augment the speed test measurement data further.

Synthesizing broadband affordability dataset (ACM SIGCOMM'23 [23]). While there are datasets on broadband availability and quality, a comprehensive dataset on broadband affordability is notably lacking. This absence hinders the ability to assess affordability, recognize discriminatory pricing practices by ISPs, and determine the impact of policy interventions like ACP, CAF, BEAD, etc.

To fill this gap, we deployed BQT to compile the first extensive dataset on broadband affordability, covering over 1 million street addresses in thirty cities serviced by seven major ISPs. Using this dataset, we provided critical policy insights, including empirical evidence on the prevalence of monopolistic practices in quasi-competitive markets, i.e., when cable-based ISPs (e.g., Cox) are competing with DSL-based providers. The study also demonstrated how average incomes in a region influence ISPs' fiber deployment strategies.

Developing a framework to assess high-stake broadband policies (ACM SIGCOMM'24 [22]). We employed BQT to develop a post hoc evaluation framework to assess the efficacy of disparate broadband policy interventions. As a first step, we focus on recently concluded FCC's *Connect America Fund (CAF)*. This program allocates funds to internet access providers, such as Comcast, Verizon, or AT&T, to establish internet access in millions of underserved (typically rural) locations. We investigate whether these addresses indeed have internet access and how the quality of this access compares to that of nearby addresses or those in urban centers.

Our analysis reveals significant discrepancies between ISP-reported data and actual broadband availability. More concretely, we find the compliance rate—defined as the weighted fraction of addresses where ISPs actively serve and advertise download speeds above the FCC's 10 Mbps threshold—is only 33.03% for this multi-billion dollar program. The implications of this effort could be profound: It might empower policymakers to (1) confront ISPs that fail to fulfill their obligations to the FCC, (2) enhance post-hoc verification of ISPs' claims, (3) assess the quality of service offered by these funded ISPs (where they do provide service), and (4) potentially re-evaluate the program in its entirety, possibly redirecting funds to more effective methods of improving broadband internet access.

Broader impact. Though academically, this line of work resulted in the **ACM SIGCOMM IMC Distinguished (Long) Paper Award** in 2022 and the **ACM SIGCOMM Dissertation Award** in 2024, its impact resonates beyond academic circles. Specifically, in response to our findings, (1) the FCC now treats DSL-served locations as unserved, (2) the FCC updated policies on the use of crowdsourced speed test data points in the BEAD challenge process, necessitating the reporting of subscription tiers alongside speed

test data points, and (3) the city of Los Angeles authorize the Civil, Human Rights, and Equity Department (CHRED) to process complaints related to digital discrimination. Moreover, since we published our findings on this topic, we have been receiving requests from a diverse group of nonprofits to help them curate datasets using BQT to support their advocacy efforts. For instance, we are currently assisting Education Superhighway in its efforts to highlight the poor state of broadband offerings in multi-dwelling units in underprivileged urban areas.

More recently, our research findings uncovered potential fraudulent activity by various Internet Service Providers (ISPs) that have benefited from the Connect America Fund (CAF) program [22]. In response to these findings, we are currently assisting the Colorado Attorney General (AG) in an ongoing investigation to gather additional evidence of potential fraud. Furthermore, our team is in the process of preparing an amicus curiae brief for submission to the United States Supreme Court in relation to an ongoing False Claims Act (FCA) case against Wisconsin Bell, a subsidiary of AT&T, involving alleged fraudulent activity under the E-Rate program.

I am also proud of my role in cultivating a community committed to an equitable digital future. Specifically, I organized the "Bridging the Divide" workshop in June 2023. This event stood out for bringing together a diverse group of experts and stakeholders, including government policymakers, researchers from fields like Internet studies, law, and social sciences, as well as representatives from non-profit organizations. The keynote speaker was **former FCC Commissioner Mignon Clyburn**, and participants included officials from the cities of Los Angeles, Oakland, and Santa Cruz, along with representatives from the states of Illinois and California. Federal agencies such as the FCC and NTIA were also present, focusing on the need for better data and identifying gaps in current policies. The workshop was instrumental in shaping a research agenda aimed at using data to inform policy decisions—an area that now plays a central role in my current and future research efforts. One of the attendees for this workshop commented on the workshop, "*I attended the 2023 NSF Workshop on "Bridging the Divide: Answering Internet Policy Questions with Cutting-Edge Network Measurement Algorithms, Datasets, and Platforms," a workshop that Arpit co-organized. This workshop was, by far, the best workshop I have ever attended, with an extraordinary emphasis on impact and in-depth conversations. I do not say the following lightly: the workshop changed how I think about my work and my life.*"

2 Developing "Production-Ready" AI and ML Artifacts for Networking.

As I started my journey at UCSB, I became intrigued by the transformative effects of AI and ML in other application domains, such as vision and natural language processing. I co-organized a series of workshops [2,3] to explore if and how AI and ML could be harnessed for networking, with the ultimate goal of developing self-driving networks capable of managing complex, performant, and secure networks with limited resources and expertise.

In my journey, I realized that learning is inherent to networking, and most research in networking aims to infer the unknown, be it the bottleneck link capacity and latency for TCP, throughput for adaptive bit rate selection, or attack signatures for intrusion detection. We have been solving these problems using simpler rule-based heuristics, but recent works, including some of my own, demonstrated that ML-based solutions could outperform existing heuristics.

However, I soon recognized that network operators were hesitant to deploy or even test these ML artifacts in production environments. Over time, it became clear that this reluctance stemmed from a lack of trust in these models where many operators had experienced model generalizability issues in the past, i.e., they observed that existing ML artifacts do not perform as expected in real-world scenarios. I noticed that model developers, who had uncritically adopted the standard ML pipeline common in other domains, struggled to determine (1) what constitutes the "right data" for a specific learning problem in a given network environment, (2) how and from where to collect or access this "right" data, and (3) how to ascertain if the model is making correct decisions, and if, how, and when it might fail. As a result, they were developing ML artifacts that lacked trustworthiness for production deployment.

Given these observations, my research aims to bridge the trust gap in "ML for networks" by advocating for and developing an innovative **closed-loop ML pipeline** that challenges the conventional standard ML pipeline. In contrast to the traditional approach, which predominantly focuses on optimizing the model's performance and necessitates the upfront selection of the "right" data, our newly proposed ML pipeline prioritizes an iterative approach to finding the "right" for a target learning problem and network environment.

My research so far has focused on developing the key building blocks, Trustee [19], netUnicorn [8], and PINOT [7], for realizing the proposed closed-loop ML pipeline for networking.

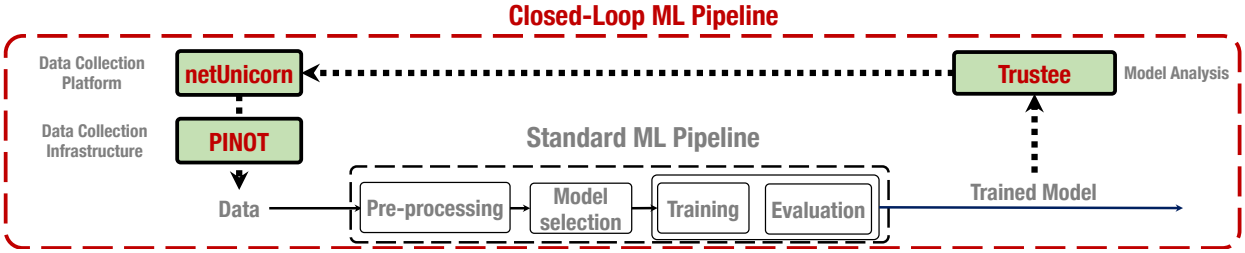

Figure 1: ***Building blocks for closed-loop ML (for networks) pipeline.***

Trustee: A model explainability framework (ACM CCS'22 [19]). When I forayed into "ML for networking" and sensed a reluctance to ML-based solutions among network operators, I realized that most people had internalized that an ML model won't work in production settings, but there was no principled approach to demonstrate so or reason why that's the case.

Given these observations, developing a model-explainability tool that cracks open the black-box ML model to explain its decision-making. Being able to analyze how a model makes its decision would help identify whether a trained ML model is underspecified and which (if any) undesired inductive biases (e.g., shortcut learning) are responsible for it. However, we realized that we cannot simply leverage existing tools for this purpose. Most existing model explainability tools only offer local explainability for specific data points. Such local explanations are not suited to identify underspecification issues. The ones that do offer global explainability only work for a specific model class.

To bridge this gap, we developed Trustee, a global explainability tool that reveals the decision-making process of black-box ML models for networking. By utilizing the trained black-box model and the associated training data, Trustee synthesizes a high-fidelity, low-complexity, and stable decision tree that explicates the decision-making process for the majority of data points, providing domain experts with valuable insights to identify underspecification issues.

We used Trustee to analyze the decision-making of various publicly available state-of-the-art ML artifacts, especially the ones developed for network security problems. Through our analysis, we demonstrated the prevalence of underspecification issues—offering a more principled approach to explore and communicate why a model won't generalize in production settings. Our efforts in this space have been recognized through the **IRTF/IETF Applied Networking Research Prize (2023)** and **ACM SIGSAC CCS Best Paper Honorable Mention (2022)**.

netUnicorn: A flexible and extensible data-collection platform (ACM CCS'23 [8]). Though Trustee did a good job articulating what's wrong with a given ML model, it offered no solution on how to fix the identified underspecification problem. The netUnicorn work focuses on fixing the model's vulnerability to underspecification issues. Since all the problems that the Trustee identified were attributable to data quality issues, it specifically focuses on curating the "right" data.

While the networking community has made many efforts to collect (labeled) network data for a variety of learning problems, most of these efforts are inherently "fragmented". That is, they have invested in developing over-specialized solutions that could only collect data for a specific learning problem from a specific network environment. Furthermore, these efforts often treat data collection as a one-time task, unrealistically assuming that researchers know in advance what the "right" data is for a given learning problem and how to curate it. This approach offers little to no flexibility to modify the data collection process. These limitations affect the ability to curate high-quality labeled data, which, in turn, affects the ML model's ability to generalize as expected in target settings.

To address this fundamental limitation that has ailed the networking community for decades and simplify network data collection, our team developed netUnicorn [8], a novel data-collection platform that effectively realizes the "thin waist" of the classic hourglass model. Here, different learning problems comprise the top layer, and the different network environments constitute the bottom layer.

netUnicorn realizes the "thin waist" by developing novel disaggregated programming abstractions for data collection. Specifically, it decouples data-collection intents from deployment mechanisms. Consequently, users no longer need to concern themselves with how to realize an intent or manage various runtime issues. Moreover, netUnicorn decomposes the data-collection intents into independent and reusable tasks. These

can be combined to form various data-collection pipelines. This flexibility not only simplifies expressing data-collection intents for diverse learning problems but also supports the iterative refinement of intents for the same problem—a crucial aspect of the proposed closed-loop approach. These pipelines can be applied to any set of data-collection hosts, facilitating the emulation of different network environments using one or more network infrastructures (e.g., PINOT, AWS, Azure). Such design choices ensure that netUnicorn satisfies the high-level data-collection intents with high fidelity with minimal computing and communication overheads.

It is worth noting that the combination of Trustee and netUnicorn lays the foundation for the desired closed-loop workflow. Trustee simplifies the process of identifying problematic data skews that manifest as underspecification issues, while netUnicorn facilitates the iterative refinement of data collection intents to selectively address these skews selectively, thereby providing better opportunities for developing generalizable ML models.

PINOT: A programmable data-collection infrastructure for campus/enterprise networks (ACM ANRW'23 [7]). An important question that many academic researchers and network operators have to answer while developing ML-based solutions for networking is where will I get the data for training. One of my position papers [6] on this topic alluded to the prevailing "data divide" where only a select few entities (e.g., Google, Microsoft, etc.) have access and resources to collect the "right" data required for developing production-ready ML artifacts. Consequently, most researchers and practitioners are forced to work on a limited set of learning problems for which some data is publicly available or which aligns with data owners' (narrow) interests. Both these situations are a big impedance to the development of production-ready generalizable ML models for networking problems.

To address this fundamental problem, I laid the roadmap to democratize networking research in the era of AI and ML. Specifically, I argued that networking researchers should consider transforming their campus network into a programmable data-collection infrastructure. To this end, my team developed PINOT [7], which comprises a suite of active and passive data-collection tools designed to facilitate the transformation of campus and enterprise networks into scalable data-collection infrastructures. Such a transformation offers researchers access to high-quality labeled data from diverse environments emulated using a production physical network. We have instantiated (a subset of) this roadmap at UCSB, which entails deploying a network of hundreds of programmable single-board computers (e.g., RasPis) and programmable switches at various strategic locations (e.g., border gateways, dorms, library, etc.). We have been using this infrastructure to collect (labeled) network data for different learning problems from UCSB's production network.

Broader impact. Though my efforts so far have not yet completely solved the challenge of developing production-ready ML artifacts for networking, they establish foundational elements for a much-needed closed-loop ML pipeline. Since it is unfathomable that a single group could solve such fundamental problems on its own, I have extensively focused on creating a community of like-minded researchers and practitioners to make deployable self-driving networks a reality [12].

As a first step, we made all three systems publicly available [1, 4, 5], equipped with all the required documentation and tutorials[1] to lower the engagement thresholds. Our team aided different campus networks, which include Columbia University, the University of Chicago, and the University of Oregon, to use PINOT to transform their campus networks into a programmable data-collection infrastructure. We are actively collaborating with Dave Taht at LibreQoS (deployed by 30+ wireless ISPs in the US) and ViaSat to realize closed-loop ML pipelines for different last-mile networks—enabling iterative data collection from production last-mile networks for the first time.

Due to my intellectual leadership in this critical research area, I have been collaborating with multiple industry partners in various capacities. Specifically, I have worked closely with Walter Willinger, Chief Scientist at the cybersecurity company NIKSUN Inc., on several key projects. I am also a regular panelist at Google's annual Networking Research Summit, where I provide insights on disruptive trends in networking, such as the increasing role of AI and machine learning. Additionally, I have been working with research teams at Google, Amazon Web Services, and ViaSat to develop production-ready machine-learning solutions for various network-related challenges. I also provide consulting services to Beegol Inc., a Brazil-based networking startup that delivers machine learning-based network diagnostics for Internet service providers. Most recently, I was appointed Faculty Scientist at Lawrence Berkeley National Laboratory (LBNL), where

[1] ACM SIGCOMM 2023 Tutorial: Closed-Loop "ML for Networks" Pipelines

I am contributing to AI-driven network operations at Energy Sciences Network (ESnet), which provides critical connectivity for all Department of Energy (DoE) scientific facilities.

3 The Road Ahead

Ongoing Efforts

To date, my work has laid the groundwork for combating digital inequity, developing production-ready AI/ML artifacts for networking, and facilitating effective data-driven policymaking. However, there is a need to build on this foundation and turn these initial strides into more substantial achievements, leading towards a digitally equitable future. In the short term, I aim to enhance my contributions to data-driven policymaking and the development of production-ready self-driving networks.

Developing foundation models for networking. My primary focus has been on addressing the fragmentation in data-collection efforts in the "ML for networks" domain. This fragmentation extends beyond data collection to the representation of network data and the development of learning models. Over the past two decades, the development of ML models for networking has concentrated on identifying specific features and model architectures tailored for particular learning classes and environments. This approach has resulted in models with limited adaptability to new problems and environments. Moreover, these models are often trained on sparse and noisy labeled data—collected from unrealistic network settings, failing to utilize the abundance of unlabeled network data from production networks.

In response to these challenges, we have explored the possibility of developing a holistic representation for network data that efficiently uses abundant unlabeled network data and can be easily adapted to various learning problems and environments. Our efforts culminated in the development of **netFound**, a foundational model for networks. netFound is designed from scratch to effectively leverage networking data's unique characteristics, such as protocol semantics, inherent hierarchy, and multi-modality. It uses unlabeled packet traces from UCSB for pre-training and employs various labeled datasets curated via netUnicorn for fine-tuning. The model's modular design enables it to tackle a broad spectrum of learning problems across different spatial and temporal scales—addressing the fragmentation issue. Our results are promising, demonstrating that netFound can autonomously learn the semantics of TCP and significantly enhance the generalizability of ML artifacts for various downstream tasks. Recently, I received in-kind support from the Department of Energy (DoE), allowing me to use their high-performance computing resources (100k+ GPU hours/year) to train production-ready netFound, exposing the model to trillions of tokens while leveraging the newly developed closed-loop ML pipeline.

Our current endeavors mark only the initial steps in fully realizing the potential of this approach. Unanswered questions include how to decipher the model's decision-making processes, tackle underspecification issues, identify the right data for pre-training or fine-tuning, and extend netFound to more complex networked systems. Successfully resolving these questions could revolutionize our approach to current learning challenges and enable us to address more complex learning problems that are beyond the reach of current tools.

Developing novel broadband-quality assessment tools. In recent years, there have been significant improvements in the curation of high-quality data to assess broadband availability (e.g., the National Broadband Map) and affordability (e.g., Broadband Affordability Datasets curated by our broadband-plan querying tool, BQT). However, high-fidelity broadband quality data at a national scale is still lacking. Currently, broadband quality assessments are based on offered speed tiers or aggregate quality of service (QoS) metrics (e.g., upload/download speed, latency, jitter, etc.) reported by disparate crowdsourced speedtest measurement tools. Unfortunately, none of these metrics truly reflect the network quality experienced by end users. This disconnect is due to the fundamental challenge of measuring network quality, especially under conditions of dynamically varying congestion. Moreover, this mismatch contributes to a relatively sparse measurement footprint, affecting the representativeness of the broadband quality datasets available to policymakers.

I am exploring how to develop of a new network measurement framework that could simplify the collection of high-fidelity broadband quality datasets at scale—bridging the gap between reported and experienced network quality. To realize this goal, I plan to: (1) Develop an attention-based network foundation model that could learn dynamic network representation (DNR) vectors from packet traces of active measurement tools (e.g., speedtest). These traces reflect interactions between protocol semantics (e.g., TCP, speedtest) and

dynamic network conditions. Our preliminary work, netFound (described above), demonstrates the feasibility of extracting latent network context using an attention-based network foundation model and serves as the intellectual foundation for this thrust. (2) Develop novel QoS-QoE algorithms to map DNR vectors from active measurements to QoE metrics for disparate networked applications of interest (e.g., YouTube, Zoom, etc.). The preliminary work of developing netReplica, a data-collection platform that simplifies collecting network data for disparate control applications (e.g., speedtest, YouTube, Zoom, etc.) under the same set of realistic dynamic network conditions, enables the curation of the QoE dataset required for this thrust. (3) Develop context inference algorithms that leverage the DNR vectors to learn hidden and hard-to-measure networking contexts, such as last-hop medium (wired vs. wireless), location of the bottleneck link (home vs. last-mile network vs. peering links, etc.)—offering opportunities to diagnose QoE degradation events. netReplica offers a unique opportunity to curate labeled datasets for different context inference problems in this thrust.

I am hoping that this effort would fundamentally change how we measure networks and report QoS/QoE metrics. Being able to report broadband quality using metrics that truly reflect users' experiences would empower end users to make better choices. More importantly, it would enable the large-scale curation of high-fidelity broadband quality datasets, empowering policymakers to accurately assess the state of broadband quality in a region—both before and after multi-billion-dollar interventions, infrastructure investments, and other programs.

Longer-term Plans

I view my contributions and achievements to date as a first step towards realizing my vision of a more equitable digital future. To achieve a truly equitable digital future, it will be essential to develop sophisticated **socio-technical systems** that build on our existing foundations. These systems will play a crucial role in driving innovations to overcome the key technical challenges hindering digital equity. My aim is to foster a diverse ecosystem that not only expedites the development of these systems but also leverages technological progress to drive societal change. By aligning technological capabilities with societal needs, this ecosystem is designed to ensure that advancements in digital infrastructure and policymaking translate into tangible benefits for all communities. This approach sets the stage for a future where digital equity is achievable worldwide. In the near future, I plan to focus on two such socio-technical systems.

Open and sustainable digital infrastructures for data-driven policymaking. I recently organized a workshop that brought together various experts dedicated to the cause of digital equity. The participants highlighted that although numerous tools and datasets exist, they are often scattered, not readily accessible, and vary in quality. These discussions motivated the necessity for an open and sustainable digital infrastructure that would centralize data resources and spearhead collaborative projects to both analyze and expand current and new data sets.

Aligning with these insights, I plan to leverage the existing building blocks to develop a digital infrastructure that amalgamates and eases access to broadband-related datasets for research and policy development. This infrastructure aims to generate solid empirical evidence that will (1) assist policymakers in making informed decisions, (2) be independent of potentially flawed or biased self-reported ISP data, and (3) enable marginalized communities and advocacy organizations to pursue digital equity at both the federal and local levels. In this endeavor, I intend to utilize my expertise in machine learning for networks to expand the collection of high-quality and longitudinal broadband data. For instance, I aim to employ netFound to contextualize existing speed test measurements, reduce measurement overhead, and implement reactive network measurements to optimize the use of available measurement infrastructure.

Community-wide infrastructure for self-driving networks. I convened another workshop that highlighted the lack of a framework that shares domain knowledge across networks, creates and tests generalizable ML models before deployment, and enhances the accurate identification and mitigation of network events while also promoting data ownership, preserving privacy, and enabling collaborative learning and independent validation of each others' findings.

To this end, I plan to develop a community-wide infrastructure with four main goals: (1) facilitate flexible and high-quality data collection efforts that are easy to replicate in different networks, (2) provide an innovative framework for collaborative data labeling and privacy-preserving knowledge sharing, (3) design a principled approach to developing generalizable learning models for networking problems by proposing a

new "closed-loop" ML pipeline, and (4) establish a pathway for deploying NetAI solutions via novel practical sandboxing capabilities for road-testing solutions within an individual and across different networks.

Democratize access to AI and ML. I have primarily viewed AI and ML as tools to reduce the operational costs of maintaining performant and secure networks, thereby addressing digital inequity. However, the scope of AI and ML extends well beyond these traditional applications. In the future, societal implications of unequal access to AI and ML tools in various fields (e.g., precision medicine, large-language models) could become as critical a topic as broadband connectivity is today.

The trend towards centralizing AI and ML technologies is a matter of grave concern, especially in domains with significant societal impact, like precision medicine and large-language models (LLMs). In precision medicine, AI and ML can lead to groundbreaking personalized health interventions, such as custom treatments and early disease detection. LLMs, on the other hand, promise to transform our interaction with digital information, encompassing text and images. However, misuse of or unequal access to these advanced tools could lead to severe consequences, widening the gap in benefits and opportunities.

While precision medicine and LLMs are prominent examples, my broader interest lies in exploring the critical enablers of such transformative AI/ML applications. This includes identifying new data sources (like wearable sensors for precision medicine) and innovative model architectures (such as attention-based models for LLMs). Equally important is understanding and addressing the socio-technical challenges that hinder the democratization of these technologies. My focus is on identifying these barriers and developing strategies to overcome them, ensuring equitable access and application across various societally critical domains.

Policymaking for futuristic digital societies. As we advance towards a more digital future, with an increasing number of transactions becoming digitized, this progress has been notably transformative in countries like India and China. However, this rapid digitalization has also shifted significant control to the state, posing risks to personal liberties and privacy, as opposed to control by private entities. The research community, traditionally focused on accountability and privacy issues associated with private entities, now faces the need to refocus on these societally critical questions in light of these new developments. A key research question that interests me is: How can we collect the right data to characterize this shift in control over our digital lives? This extends to identifying the necessary tools and infrastructures for data collection and exploring ways to refine the quality of this data iteratively. The ultimate goal is to develop new policies that impose checks and boundaries on government management of this newly acquired digital prowess, ensuring a balanced and responsible approach to digital transformation.

4 Concluding Remarks.

As a mission-driven researcher, my journey has involved setting ambitious goals and addressing emerging problems, leading to significant contributions in my field. The initial years as a junior faculty member were especially challenging, marked by a foray into new research areas with little overlap with my prior expertise in programmable networks [9, 13, 15–18, 20, 21, 21] and network telemetry [11, 14]. This period coincided with the global pandemic, adding the complexities of navigating an unfamiliar academic environment and establishing a research group amid global disruptions.

Despite these challenges, the past two years have been fruitful. I have successfully secured approximately $4.3 million in funding for various projects (includes both continuing and new grants), personally contributing around $1.9 million. It is worth pointing out that my CAREER proposal in 2023 was desk-rejected due to the missing department chair letter in the final submission. Though I have not secured the CAREER award, I have received a single PI grant of $600k from NSF in 2023.

My research has been recognized with multiple awards, including the **Applied Networking Research Award (2023)**, **ACM CCS Best Paper Award (Honorable Mention)**[2] and the **ACM SIGCOMM IMC Distinguished Paper Award (2022)**. My first graduating student, Udit Paul, received the **ACM SIGCOMM Dissertation Award** in 2024. It is worth highlighting that these were also *my first papers at CCS and IMC, respectively*, demonstrating my approach to making fundamental contributions to various problems, irrespective of the area of specialization. These accolades complement the recognition I received at Princeton University, including the **USENIX NSDI Community Award** (2016) and others. More than these awards, I value the significant impact of my work and my role in fostering a community dedicated to an equitable digital future. Looking ahead, I am committed to continuing this mission and making foundational

[2]This was the best paper in the "ML for Security" track at CCS.

contributions to the fields related to my research.

References

[1] netunicorn: A network data-collection platform—https://netunicorn.github.io/. (Cited on page 5.)

[2] Nsf workshop on measurements for self-driving networks, 2019. (Cited on page 3.)

[3] Nsf workshop on self-driving networks, 2018. (Cited on page 3.)

[4] Pinot: A programmable infrastructure for ai/ml for networking—https://pinot.cs.ucsb.edu/. (Cited on page 5.)

[5] Trustee: A model explainability framework for Networking—https://trusteeml.github.io/. (Cited on page 5.)

[6] Arpit Gupta, C. MacStoker, and W. Willinger. An Effort to Democratize Networking Research in the Era of AI/ML. In *ACM HotNets*, 2019. (Cited on page 5.)

[7] R. Beltiukov, S. Chandrasekaran, A. Gupta, and W. Willinger. Pinot: Programmable infrastructure for networking. In *Proceedings of the Applied Networking Research Workshop*, ANRW '23, page 51–53, New York, NY, USA, 2023. Association for Computing Machinery. (Cited on pages 4 and 5.)

[8] 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 4.)

[9] R. Birkner, A. Gupta, N. Feamster, and L. Vanbever. SDX-Based Flexibility or Internet Correctness?: Pick Two! In *ACM Symposium on SDN Research (SOSR)*, 2017. (Cited on page 8.)

[10] Broadband-plan Querying Tool (BQT)-https://address.cs.ucsb.edu/. (Cited on page 1.)

[11] A. Gupta, R. Birkner, M. Canini, N. Feamster, C. Mac-Stoker, and W. Willinger. Network Monitoring as a Streaming Analytics Problem. In *ACM Workshop on Hot Topics in Networks (HotNets)*, 2016. (Cited on page 8.)

[12] A. Gupta, R. Durairajan, and W. Willinger. Special issue on the acm sigmetrics workshop on measurements for self-driving networks. *ACM SIGMETRICS Performance Evaluation Review*, 2023. (Cited on page 5.)

[13] A. Gupta, N. Feamster, and L. Vanbever. Authorizing Network Control at Software Defined Internet Exchange Points. In *ACM Symposium on SDN Research (SOSR)*, 2016. (Cited on page 8.)

[14] A. Gupta, R. Harrison, A. Pawar, M. Canini, N. Feamster, J. Rexford, and W. Willinger. Sonata: Query-Driven Network Telemetry. In *ACM SIGCOMM*, 2018. (Cited on page 8.)

[15] A. Gupta, R. MacDavid, R. Birkner, M. Canini, N. Feamster, J. Rexford, and L. Vanbever. An Industrial-Scale Software Defined Internet Exchange Point. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2016.
Community Award. (Cited on page 8.)

[16] A. Gupta, L. Vanbever, M. Shahbaz, S. P. Donovan, B. Schlinker, N. Feamster, J. Rexford, S. Shenker, R. Clark, and E. Katz-Bassett. SDX: A Software Defined Internet Exchange. In *ACM SIGCOMM*, 2014. (Cited on page 8.)

[17] R. Harrison, C. Qizhe, A. Gupta, and J. Rexford. Network-Wide Heavy Hitter Detection with Commodity Switches. In *ACM Symposium on SDN Research (SOSR)*, 2018. (Cited on page 8.)

[18] X. Hu, A. Gupta, A. Panda, N. Feamster, and S. Shenker. Preserving Privacy at IXPs. In *ACm APNet*, 2018. (Cited on page 8.)

[19] 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 4.)

[20] H. Kim, J. Reich, A. Gupta, M. Shahbaz, N. Feamster, and R. Clark. Kinetic: Verifiable Dynamic Network Control. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2015. (Cited on page 8.)

[21] R. MacDavid, R. Birkner, O. Rottenstreich, A. Gupta, N. Feamster, and J. Rexford. Concise Encoding of Flow Attributes in SDN Switches. In *ACM Symposium on SDN Research (SOSR)*, 2017.
Best Paper Award . (Cited on page 8.)

[22] 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 pages 2 and 3.)

[23] 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 2.)

[24] 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 2.)