Paul Schmitt **Research Statement**

My research passion is to design and build networked systems that are informed by empirical measurements, specifically in contexts that impact billions of people as they connect to the world. My research is rooted in deep expertise in protocols and architectures that are used at massive scales and yet are often ignored by researchers. To create impact, I take a dirty-slate approach to networked systems research, seeking to maximize compatibility and immediate deployability within current environments. The contributions of my research to-date are in two primary dimensions:

1. **System and protocol design** of solutions for providing connectivity, enhancing security and privacy, and extracting the features necessary for accurately modeling performance.
2. **Modeling and measurement** of existing systems and applications to understand inefficiencies, monitor networks, and inform designs of new networked systems and architectures.

The research questions I pursue come from careful observation of real world challenges, and I seek to address novel research questions or to reconsider conventional wisdom on existing topics. I take a holistic view of systems research, considering both technical novelty and depth as well as potential industry and policy implications that result from system design choices. My work has led to changes in systems used by industry, has generated interest from policy organizations and the popular press, and has been directly involved in commercialization efforts.

Prior Work

My previous work focused on under-resourced cellular networks, the global domain name architecture, and application and Internet measurement methods from multiple vantage points.

Characterizing and offloading under-resourced cellular networks. Millions of people throughout the world live in areas at the fringes of cellular connectivity, where the cost of augmenting infrastructure exceeds the expected return on investment. Local cellular networks have recently gained traction as a solution for providing connectivity in areas lacking coverage using small base stations. In order for local networks to augment and coexist with under-resourced commercial cellular networks, they must be able to determine if the nearby commercial infrastructure is overloaded—operational knowledge that is typically not available outside of the operator's core—and have a mechanism for offloading traffic. To accomplish these goals, we designed HybridCell [11], which includes a method that enables congestion detection at the radio layer by third-party observers. Given the presence of congestion, HybridCell dynamically offloads users proportionally with the level of congestion. Our work was not limited to networked systems design; our characterization techniques were used to analyze the cellular infrastructure serving Za'atari, a large refugee camp in Jordan. In our study we found that most camp-serving infrastructure was under-provisioned, and that the quality of cellular coverage and data speeds varied drastically between different locations as well as between the different cellular carriers in the area. Meaning that, even within the confines of a single camp, there were substantial community-level divides [10]. My work in this area has led to additional collaborations with aid agencies (*e.g.*, USAID, UNHCR) that serve refugee and displaced populations, and the methods continue to be used to characterize cellular networks in fringe areas [8].

Enhancing DNS privacy. Virtually all Internet services depend upon the Domain Name System (DNS) to resolve human-readable webpage names (URLs) to IP addresses. Unfortunately, the protocol as it was originally designed can reveal sensitive information such as the website a user visits or the devices they own. As a result, new privacy-focused DNS protocols have been developed that obscure a user's DNS queries from his or her Internet service provider. Yet, these systems merely transfer trust to another third party. In our work on Oblivious DNS (ODNS) [9], we argued that no single party ought to be able to associate DNS queries with a client IP address that issues those queries. ODNS breaks the link between a user's identity (*i.e.*, IP address) and their DNS traffic at a recursive resolver. ODNS adds a layer of obfuscation by operating its own authoritative namespace; the authoritative servers for the ODNS namespace act as recursive resolvers for the queries that they receive, but they never see the IP addresses for the clients that initiated these queries. Critically, ODNS is backwards-compatible with existing infrastructure, and through experimentation we have found that it introduces minimal performance overhead, both for individual queries and for web page loads. ODNS has led to multiple IETF[1] drafts, both from industry and from academia, and its underlying techniques have been adopted in an implementation by Cloudflare, Apple, and Fastly, potentially enhancing privacy for hundreds of millions of users moving forward[2].

[1] The Internet Engineering Task Force (IETF) is the standards body that develops and promotes standards for Internet protocols.

[2] `https://blog.cloudflare.com/oblivious-dns/`

Application performance inference from encrypted traffic. Internet service providers require insight into the performance of applications on their networks in order to manage them appropriately. In the past, ISPs could simply inspect application traffic as it was unencrypted; however, that is no longer the case with most traffic. Video streaming is a particularly important application as it is responsible for a large volume of traffic on most ISPs and the user experience can be greatly impacted due to poor network conditions. In prior work, we developed machine learning models that infer quality metrics (*i.e.*, startup delay and resolution) for encrypted streaming video services [3]. In our work, we developed techniques for identifying application-layer features (*i.e.*, video sessions, video segments) from a mix of traffic. We also developed a model that is applicable across multiple video streaming services, each with potentially different goals and characteristics (*i.e.*, Netflix may tune their ABR algorithms to avoid rebuffering, whereas YouTube may tune theirs to begin playback as quickly as possible). We demonstrated the potential for such models through a 16-month deployment in 66 homes and provided new insights about the relationships between Internet "speed" and the quality of the corresponding video streams; we find that higher speeds provide only minimal improvements to startup delay and resolution. This work brought attention from outside of academia, resulting in an article in the Wall Street Journal[3]. Our models are also now being included in commercialization efforts.

Internet performance and infrastructure measurement. In order to improve systems, protocols, and applications, we must understand how they perform and any inefficiencies they may contain. In my work I have measured several aspects of the Internet, including user-facing protocol performance and impacts of core network architectures on traffic.

As the DNS protocol is fundamental to nearly all Internet traffic, I have worked to measure the performance of different variants [5, 6]. We found that, while encrypted DNS protocols add some overhead, they can outperform conventional DNS in many cases when measuring overall page load times due to the inherent benefits of the TCP transport compared with UDP. However, as network conditions deteriorate, conventional DNS may be a more optimal choice. The overall suggestion in this work is that no protocol is always best, and that users may be best served by a system that allows for DNS protocols to be selected dynamically. Our work studying the shift toward DNS over HTTPS (DoH) also resulted in a policy-focused publication [1].

Decisions that providers make while designing their core networks can have significant impact on the performance that their users experience. In my research I have studied both cellular and terrestrial ISP cores. In my study of cellular data networks [13], I discovered structural differences between mobile virtual network operators (MVNOs) and traditional mobile network operators (MNOs), even though the underlying infrastructure is often shared. My work showed that MVNO traffic is frequently sent over inefficient paths and ingresses onto the Internet in different, often suboptimal, locations. In terrestrial networks, I have recently studied ISP responses to the COVID-19 pandemic in the United States [7]. We found that the shifts in usage patterns resulted in poor latencies for a brief time around stay-at-home-orders, but eased over time. This easing coincided with service providers that aggressively augmented their core (*e.g.*, peering link) capacities at higher rates (2X) than normal. Both of these studies illustrate the usefulness of measuring provider infrastructure, as we can use such work to uncover inefficiencies in order to rectify them or to learn what works and what doesn't as requirements change, allowing providers to learn best practices for the future.

Current and Future Research Directions

Moving forward, I intend to continue building upon my expertise in networked systems design, and meeting inherent challenges as we increasingly apply modern machine learning techniques to network traffic analyses.

Systems and Protocols

In my systems work, I will continue building upon my work in cellular networks and the DNS architecture. In cellular, I look to offer users privacy where there has been none in the past, and, moving forward, to exploit the industry-wide shift toward software-based cellular cores in order to allow for novel systems research. I also aim to fundamentally re-think the global name resolution architecture, searching for ways to maximize both privacy and performance.

Leveraging modern cellular (4G, 5G and beyond) architectures for improved privacy and performance. In prior work I studied cellular network performance to identify under-provisioned infrastructure. However, this work did not address privacy concerns surrounding cellular. In recent years we have seen the seemingly-legal and massive-scale revelation, sale, and leaks—all via network operators—of hundreds of millions of users' mobile location data and call metadata due to the confluence of a deregulated industry, high mobile use, and the proliferation of data brokers. Nearly every mobile user can be physically located, along with their call metadata, by anyone with an interest and a few dollars to spend. At the core of cellular privacy is a conundrum: the cellular architecture[4] requires operators to "know"

[3] "The Truth About Faster Internet: It's Not Worth It". Wall Street Journal. Front Page (A1) Feature. August 2019.

[4] Including 5G, which improves privacy in several dimensions yet does not remedy the fundamental issue.

all user locations and identities in order to provide connectivity. We are developing Pretty Good Phone Privacy [12] (PGPP), which changes the cellular core architecture such that network operators need not maintain sensitive information about their users such as individual identity and location. PGPP leverages points of decoupling in the architecture to separate connectivity from authentication and billing, allowing us to alter the identifiers used to gain connectivity and enable operators to authenticate and bill users without identifying them. Our approach is made feasible by the industry-wide shift toward software-based cellular cores. In our initial work we have found that PGPP significantly increases anonymity where there is none today and the control traffic overheads caused by our architectural changes can be kept to a manageable level with simple changes. This work opens the door to a new type of cellular operator, which we call Private Mobile Virtual Network Operators (PMVNOs), that will provide privacy-preserving cellular service that is otherwise indistinguishable from other providers. Moving forward, the flexible (*i.e.*, software-based) nature of modern cellular networks has significant potential for mobile systems research—not limited to privacy-focused work—compared with previous cellular generations as the technologies were more arcane and diverse and the knowledge was more siloed. In the coming years I will pursue research focused on the potential for shifting computation and functionality that currently occurs in the cellular core or in phones to the network edge (*i.e.*, telecom central offices), leveraging advances in network function virtualization (NFV) and programmable network hardware to promote edge compute. Additionally, most mobile testbeds require researchers to have significant background in cellular architecture, which often takes years to master. As a professor, I intend to create a cellular testbed that abstracts away the cellular-specific components of the network, enabling non-cellular researchers to study and design novel systems for future mobile networks.

Re-architecting name resolution. In recent years there has been a push toward privacy-focused DNS protocols where DNS is tunneled over encrypted transport such as TLS (DNS over TLS, or DoT); or, more recently, HTTPS (DNS over HTTPS, or DoH). Additionally, whereas DNS configuration was done at the operating system level in the past, applications are now individually choosing their own DNS protocols and configurations. While these modern DNS protocols improve privacy from eavesdroppers, this shift in the landscape introduces concerns surrounding the centralization of queries to the operator chosen by applications, typically to large cloud providers. From a privacy perspective, the cloud providers are able to collect even more data about users than before. There are also performance concerns: some CDNs use client localization techniques based on the location of the client's DNS resolver. Providers that operate resolvers have the ability to make this type of client localization more challenging, meaning that these CDNs (perhaps the resolver's competitors) would face significant operational costs merely to stay competitive [1]. In the coming years, I will explore possible new architectures and solutions for name resolution, working to create mechanisms that allow the current and future tussles around themes such as centralization, management, and content delivery to play out between stakeholders. I will also explore performance, privacy, and policy implications for systems based on new architectures.

Modeling and Measurement

Advances in ML allow for unprecedented network and application characterization. I am interested in working on problems related to 1) gathering network data in a way that is useful for ML, such as feature extraction and traffic representation; and 2) model drift detection and retraining.

Cost-aware network traffic feature representations for machine learning pipelines. Network management tasks increasingly rely on ML models to classify traffic by type or identify important events from network traffic. The performance (*i.e.*, accuracy) for any model is affected as much by the chosen features of its input as it is by the choice of model or its parameters. Additionally, systems-related costs must be considered when features are designed and selected in order to determine whether the features could feasibly be captured at scale and at speed. For example, some metrics can easily be computed, whereas others (*e.g.*, end-to-end latency), require maintaining per-flow state and observations of bidirectional flows—which is often heavyweight and potentially impossible at some vantage points.

Ideally, relationships between systems costs and model performance would be explored simultaneously when designing ML pipelines; yet, most existing network traffic representation decisions are made *a priori*, without concern for future use by models. To enable this exploration, we have created Network Microscope [2], a system designed to offer flexibly extensible network data representations, the ability to assess the systems-related costs of these representations, and the effects of different representations on inference model performance. Our initial findings using Network Microscope to evaluate existing machine learning models while differing input features show that there is no universally ideal data representation for all inference tasks, and thus there is value in exploring the space. Network Microscope lays the groundwork for new directions in applying machine learning to network traffic modeling and prediction problems, allowing operators to design machine learning pipelines that both provide accurate inferences, but are also capable of handling high traffic loads. I look to extend this research by examining inherent tradeoffs in different on-path vantage points and answering questions such as: Are certain features only available in one location on the path? Can lightweight features be extracted using programmable network devices? How do features collected at different vantage points impact the inference model performance?

Generalizing network traffic representations for use by diverse models. A significant challenge when applying machine learning techniques to network traffic is the nature of how traffic features are represented. Often, significant effort is spent on manual feature engineering to arrive at a data representation that is appropriate for a given inference task. Such feature exploration and engineering is painstaking work, even with expert domain knowledge. Additionally, such manual analysis may mistakenly focus on unimportant features or omit features that either were not immediately apparent or involve complex relationships (*e.g.*, non-linear relationships between features). To remedy this, we are developing nPrint [4], a standard, packet-based representation of network traffic meant to define a generalized input for a broad range of machine learning models that does not require manual feature engineering. nPrint encodes each packet in an inherently normalized, binary representation. Our design enables machine learning models to automatically discover important parts of packets for each distinct classification task. In our initial evaluations we have found that models trained using nPrint achieve higher performance than the state-of-the-art tools for tasks ranging from operating system detection, device fingerprinting, and application identification. This work demonstrates that generalized representations can be successfully used across a broad range of tasks. Moving forward, we intend to transition from simply representing traffic on a per-packet basis, and moving to create general representations for network traffic at the flow level; and, eventually, at the application level (*i.e.*, multiple, logically-related flows). Given such information I will study fairness (mis)behavior at the application level and whether it can be enforced from lower layers of the network stack, as traditional fairness mechanisms (*e.g.*, TCP congestion control) largely operate under the outdated assumption that each application uses a single network flow.

Automating drift detection and model retraining. Over time, the assumptions behind machine learning models may no longer hold. When applying ML to network monitoring tasks, this may mean that the underlying characteristics of the network traffic may shift, meaning the models must be re-trained in order to continue being effective. Network operators need ways of determining when models become inaccurate, and the capability to distinguish model inaccuracy from problems that are inherent in the network. In the future I will work to design a framework for detecting model drift in order to appropriately schedule model retraining as a part of normal network operations tasks. Such research could lead to significant changes in the way large networks are managed and monitored, removing human intuition and bringing a more systemic approach to network forecasting.

References

[1] K. Borgolte, T. Chattopadhyay, N. Feamster, M. Kshirsagar, J. Holland, A. Hounsel, and P. Schmitt. How DNS over HTTPS is Reshaping Privacy, Performance, and Policy in the Internet Ecosystem. In *Research Conference on Communications, Information and Internet Policy (TPRC)*, Washington, DC, Sept. 2019.

[2] F. Bronzino, P. Schmitt, S. Ayoubi, H. Kim, R. Teixeira, and N. Feamster. Beyond accuracy: Cost-aware data representation exploration for network traffic model performance. *ArXiv e-prints*, 2010.14605.

[3] F. Bronzino, P. Schmitt, S. Ayoubi, G. Martins, R. Teixeira, and N. Feamster. Inferring Streaming Video Quality from Encrypted Traffic: Practical Models and Deployment Experience. In *ACM SIGMETRICS*, Boston, Massachusetts, USA, June 2020.

[4] J. Holland, P. Schmitt, N. Feamster, and P. Mittal. nPrint: Standard packet-level network traffic analysis. *ArXiv e-prints*, 2008.02695.

[5] A. Hounsel, K. Borgolte, P. Schmitt, J. Holland, and N. Feamster. Comparing the Effects of DNS, DoT, and DoH on Web Performance. In *The Web Conference (WWW)*, Taipei, Taiwan, Apr. 2020.

[6] A. Hounsel, P. Schmitt, K. Borgolte, and N. Feamster. Can encrypted DNS be fast? In *Passive and Active Measurement Conference (PAM)*, Brandenburg, Germany, Mar. 2021.

[7] S. Liu, P. Schmitt, F. Bronzino, and N. Feamster. Characterizing service provider response to the COVID-19 pandemic in the United States. In *Passive and Active Measurement Conference (PAM)*, Brandenburg, Germany, Mar. 2021.

[8] C. Maitland, R. Caneba, P. Schmitt, and T. Koutsky. A Cellular Network Radio Access Performance Measurement System: Results from a Ugandan Refugee Settlements Field Trial. In *Research Conference on Communications, Information and Internet Policy (TPRC)*, Washington, DC, Sept. 2018.

[9] P. Schmitt, A. Edmundson, and N. Feamster. Oblivious DNS: Practical Privacy for DNS Queries. In *Symposium on Privacy Enhancing Technologies (PETS)*, Stockholm, Sweden, July 2019.

[10] P. Schmitt, D. Iland, E. Belding, B. Tomaszewski, Y. Xu, and C. Maitland. Community-level access divides: A refugee camp case study. In *International Conference on Information and Communications Technologies and Development (ICTD)*, Ann Arbor, Michigan, USA, June 2016.

[11] P. Schmitt, D. Iland, M. Zheleva, and E. Belding. Hybridcell: Cellular connectivity on the fringes with demand-driven local cells. In *IEEE INFOCOM*, San Francisco, California, USA, Apr. 2016.

[12] P. Schmitt and B. Raghavan. Pretty good phone privacy. *ArXiv e-prints*, 2009.09035.

[13] P. Schmitt, M. Vigil, and E. Belding. A study of MVNO data paths and performance. In *Passive and Active Measurement Conference (PAM)*, Heraklion, Crete, Greece, Mar. 2016.