Like other critical infrastructure, the Internet requires measurement to evaluate and improve the core properties of its services: security, performance, and resilience. Unlike other types of physical infrastructure, there is no blueprint for the design and implementation of the Internet networks, and the modern Internet is an economic ecosystem of autonomous networks with heterogeneous deployments. Before the Internet "escaped from the lab", it was possible to construct hand-drawn maps of the Internet networks and their interconnections. This information continues to grow more difficult to obtain, blocking attempts to understand the core properties at any layer below the end-to-end application layer. Indeed, while network operators often understand the routing and performance within their network, the prevalence of remote services and communication across third-party networks means that at least one network on nearly every end-to-end path is opaque even to a network operator.

Independent third-party measurements by entities unaffiliated with the measured networks provides the most tractable solution to the problem of understanding and improving the Internet, but the Internet was not designed to facilitate independent measurement. All current tools to gather raw data derive from the original traceroute implementation, a 35-year old hack of Internet Protocol (IP) implementations originally designed to help operators troubleshoot router misconfigurations. These tools produce nothing more than a list of numbers without encoded information (router IP addresses) between a probing source and a target, are prone to misleading information, and are difficult or impossible to generalize. Consequently, conducting any meaningful analyses of the Internet networks—discovering undesirable paths between users and applications, evaluating the resilience of Internet infrastructure and application deployments, evaluating potential deployments of new infrastructure to increase performance, etc.—requires devising ways to collect raw data, scientifically inferring context for IP addresses, and extracting the signal from the noisy data.

Answering seemingly simple Internet measurement questions typically begins by **building a complex foundation** that often outlives the original problem motivation. My doctoral research was motivated by the potential for DDoS attacks targeting network interconnections, but the tools I built to find router-level network interconnections enabled other research inside and outside the field of computer science, and provided labels for supervised machine learning. I am currently building the scaffolding for myself and others to study two pillars of the modern Internet: the commercial access networks that provide last mile connectivity, and the public clouds that dominate the Internet application landscape.

Building Measurement Foundations to Study Core Properties as the Internet Evolves

Two problems motivated my research into access networks and clouds:

1. Small infrastructure failures in wireline Internet access networks led to widespread outages in recent years. Without knowing why, we cannot examine trade-offs of alternate deployment options or assess risk of physical attacks against the infrastructure.
2. Public clouds fundamentally changed the Internet landscape by centralizing applications in a handful of networks. Performance and public policy analyses struggle to properly reflect this centralization.

Working toward solutions for each problem required building new foundational tools and techniques that I expect will outlive my solutions and facilitate advances in the field of Internet measurement.

Access Networks: Examining Large Outages and Security Threats

U.S. broadband access networks like Comcast, Spectrum, and AT&T are critical infrastructure, but small failures in those networks often lead to widespread outages. In the past year, two high profile outages highlight this phenomenon: (1) a single facility lost power after a bomb exploded on the street outside, disconnecting all AT&T wireline and wireless customers in the Nashville metropolitan area; and (2) two fiber cuts disconnected all Spectrum customers in Maine. No longer just conduits of landline telephone and cable TV, today's access networks support 5G cellular phones, hospital and financial services, and the remote work essential to the modern economy. Despite their importance, researchers, regulators, and customers know little about the modern incarnations of the access networks, and learning the causes of the cascading failures requires a better understanding of these networks. These outages also demand a clear-eyed assessment of the risks to access networks, since they indicate the presence of single points of failure in the infrastructure that could make access networks especially susceptible to physical attack.

Learning why these small failures cause large outages, and assessing the strengths and weaknesses of current deployments, first requires that I build a foundation to learn the network structure of the various U.S. access networks. The key challenges are (1) exposing the redundant facility paths available in each region; (2) inferring a mapping from IP addresses to access network facilities; (3) identifying the buildings that share a network dependency; and (4) learning the customer IP addresses connected through the different last-mile facilities. To address these challenges I use a variety of probing sources internal and external to the networks, combine publicly available data with novel measurement techniques, and heuristically refine network topology graphs in accordance with network architecture conventions. Validating the graphs with network operators led one operator to describe them as "wonderfully accurate."

The mapping effort reveals that facility-level aggregation causes the widespread outage phenomenon, allowing small failures to cascade. ISPs appear to use dedicated access networks to provide connectivity for each geographic region with a strict hierarchy of facilities within each regional network. For the Nashville outage, the map of the AT&T regional access network in Nashville reveals that the entire access network connects to the Internet exclusively through the single facility that lost power after the explosion. In Maine, where the Spectrum outages resulted from two fiber cuts, the map indicates the only two fiber runs whose simultaneous failures could cause an outage isolated to Maine without affecting the rest of Spectrum's Northeast regional access network.

I also worked with colleagues to discover that attackers can selectively target the facility responsible for connectivity in a given neighborhood or town. As proof of concept, we connected to public WiFi access points around San Diego to reveal the geographic area connected to the same facility, and we showed that an attacker can find facility locations from publicly available hazardous materials records of the fuel stored onsite for backup power. After demonstrating the feasibility of such an attack, we considered several mitigations that could increase the difficulty of a successful attack against access network facilities and infrastructure, but all would adversely affect regular network operations. I plan to continue working to mitigate some of these risks through further consultation with access network operators.

Cloud Networks: Measuring the Cloud-Centric Internet

Similar to access networks, cloud networks impact nearly every facet of society, demanding a better understanding of the paths between cloud applications and users. Cloud networks support many of the largest Internet applications—including video streaming and conferencing, business services, IoT and mobile backends, and 5G mobile networks—but they are mostly invisible to existing data

collections and introduce new measurement challenges. I am building a foundation to analyze cloud application paths by revealing the networks that cloud traffic traverses, and the geographic locations where cloud networks hand-off packets to neighboring networks.

Using virtual machines as probing sources allows me to gather raw path measurement data from inside cloud datacenters, but required that I adapt the tools I developed during my dissertation to work well for cloud networks. Cloud networks change internal WAN routes relatively frequently, and the route changes add noise to conventional traceroute probing. Rather than increase measurement speed by conducting thousands of path measurement in parallel, allowing route changes to corrupt many measured paths simultaneously, I implemented a probing strategy that rapidly conducts traceroutes with only a single measurement in flight at a time. Through a controlled experiment, I showed that this change significantly improved the signal to noise ratio of the raw data, and yielded more accurate inferences of the networks along the paths.

I also locate where traffic exits the cloud WANs to reach different users, compare the geographic distance covered inside and outside the cloud WAN, and discover suboptimal paths where traffic could exit the cloud WANs closer to the user without changing the next-hop network. Geolocating network infrastructure is a long-standing challenge without reliable solutions, so I designed a geolocation technique specifically for cloud network interconnections. Analyzing raw traceroute data, I learned that clouds frequently hand-off traffic through Internet exchange point (IXP) public peering, where many networks interconnect to each other over a switching fabric. I can recognize when a traceroute path encounters an IXP using publicly available data, and my technique uses known IXP locations to infer the locations of network border routers. To increase coverage, I also combine round-trip time measurements from dozens of cloud datacenters to identify border router IP addresses that likely reside in the same city. I found that Microsoft Azure generally hands off traffic significantly closer to the destination compared to Amazon AWS, but that even Azure appears to persistently use suboptimal exit locations for some destinations.

To better understand the rest of path outside the cloud networks, colleagues and I used supervised machine learning to to interpret the information that network operators embed in their DNS hostnames associated with router IP addresses. These hostnames often include the geographic location and the network that operates the router, but each network uses their own conventions for encoding information. We trained our supervised learning using labels derived from round-trip time measurements and the output of my dissertation techniques, and we produced conventions for extracting network operators and geographic locations from hostnames assigned by hundreds of networks.

Finally, I am using the paths I reveal from cloud datacenters to provide context for individual paths toward the cloud, helping users and network operators understand the round-trip paths between themselves and cloud resources. Path measurements toward the cloud reveal entirely different IP addresses than the paths observed from cloud datacenters, and typically any attempt to add context to such a path measurement requires extensive probing from the same source. Leveraging IP addressing conventions, I am developing techniques that use the cloud traceroutes to infer network operators and the geographic entries into the cloud WANs from any device toward the cloud.

Future Research Directions

The future appears bright for the field of Internet measurement. Along with many colleagues, I helped successfully advocate for a new National Science Foundation (NSF) program dedicated to developing new tools and techniques for measuring the Internet. The NSF program emphasizes

the growing need for foundational Internet measurement research to support government oversight and complex network management. Other U.S. government agencies want to harness the potential benefits of the Internet as an underlying communications network, and I conducted fundamental research under Defense Advanced Research Projects Agency (DARPA) contracts to better support mission-critical communication via the Internet. Large cloud providers like Amazon, Microsoft, and Google also recognize the potential for Internet measurement, building research groups to better manage their global networks.

One research goal is to build the foundation to evaluate long-standing concerns over data sovereignty for data in-transit. After an invited talk at the Euro-IX forum, the Canadian Internet Registration Authority (CIRA) approached me to help them evaluate the Canadian government's concern that a large portion of Canada's Internet traffic crosses the U.S. border in-transit—exposing the traffic to espionage, tampering, or disruption. The problem of data sovereignty in-transit is not limited to Canada, with many countries concerned about government and commercial traffic entering certain countries or state-controlled networks. Current tools do not support scientifically assessing the extent of the problem, identifying root causes, and devising actionable solutions, but I hope to improve on work by myself and others to properly study the problem through collaboration with CIRA.

I also hope to build on my access ISP and cloud network work to study the impact of end-to-end application resiliency on typical Internet users. Recent DNS failures and BGP misconfigurations highlighted the risks of centralizing critical services, but modern Internet applications are often remarkably robust to network- and host-level failures. I intend to measure resiliency successes and failures across different layers and technologies, including links, routers, access ISP facilities, DNS clusters, and application clusters. I also expect to find degraded service, such as congestion on network paths or suboptimal route changes, that end-to-end resiliency measures cannot resolve without greater insight into the network and control over part of the routing control plane. This study should produce a longitudinal analysis of quality of experience metrics across layers, services, and geographic locations.