Research Statement

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, the modern Internet has no blueprint for its design and implementation; it 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 is now more difficult to obtain, blocking attempts to understand the core properties below the end-to-end application layer. 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. Everyone—networks, industry, and policymakers—is operating in a vacuum of information.

But the Internet is not unknowable. I am uniquely suited to address the problem of Internet opacity because of my experience solving difficult problems to more accurately measure the Internet. I create macroscopic and holistic analyses of the Internet by devising new measurement techniques and synthesizing noisy raw data with novel data science approaches. The problems that motivate my research often lack the data and tools for scientific analysis, and I approach each problem by creating a foundation that can transcend its initial motivation to serve other research endeavors. For example, the tools I built to identify router-level network interconnections enabled a wealth of topology, resiliency, and even economic research, and provided labels for supervised machine learning.

I view independent third-party measurements by entities unaffiliated with the measured networks as 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 a list of numbers (router IP addresses) between a probing source and a target, with no indication of which router, network, or geolocation the IP address is associated with. Traceroute-based tools are also prone to misleading information, and their outputs are difficult or impossible to generalize.

My recent work focused on creating the measurement and data science techniques to study two pillars of the modern Internet: (1) access networks and (2) public cloud networks. To study access networks, I designed measurement and analytic techniques to create detailed maps of U.S. access network infrastructure deployments, which allowed me to identify weaknesses in those networks that leave them susceptible to widespread outages and attack. To study public cloud networks, I received a National Science Foundation grant to create techniques that illuminate paths between public cloud applications and users, providing visibility into this increasingly opaque segment of today's cloud-centric Internet.

Evaluating the Resiliency and Security of Access Network Infrastructure

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 highlighted this phenomenon: (1) all AT&T wireline and wireless customers in the Nashville metropolitan area were disconnected when a single facility lost power after a bomb exploded on the street outside; 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 networks, 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 root causes of the cascading failures requires a better understanding of these networks. Such cascading outages also indicate the presence of single points of failure in the infrastructure that could make access networks especially susceptible to attack.

To learn why these small failures cause large outages, I built a measurement foundation to discover the network structure of the various U.S. access networks. The key challenges were (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 found probing sources capable of revealing paths through the networks, combined publicly available data with novel measurement techniques, and heuristically refined network topology graphs in accordance with network architecture conventions. Validating the graphs with my contacts at access networks led one operator to describe them as "wonderfully accurate."

Building on that work, I conducted a macroscopic study of outages in U.S. access networks to provide greater insight into infrastructure failures. For practical reasons, I could not induce infrastructure failures, so I approached the problem by looking for sources of data that could shed light on past and future outages. As part of a separate research effort, a colleague was conducting pings to nearly every access network customer. By combining that reachability data with archived path measurements and my inferred maps, I discovered dozens of widespread outages across the U.S., and learned the scale and duration of the motivating outages in Nashville and Maine.

This foundation provided new insight into the access network infrastructure, and I discovered that heavy use of facility-level aggregation causes the widespread outage phenomenon, allowing small failures to cascade. We confirmed empirically that Internet service providers 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 maps I created for the AT&T regional access network in Nashville revealed that the entire access network connects to the Internet exclusively through the single facility that lost power after the explosion. In Maine, combining the outage data with the Spectrum access network maps indicated that the outage solely affected the portion of the network in Maine. The Maine outages resulted from two fiber cuts, and I used the maps to identify 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.

After establishing that access network design introduces resiliency weaknesses, I hypothesized that an attacker could disconnect an access network facility to selectively target a neighborhood or town. I started my investigation through discussions with access network operators, learning that concerns of manageability, recovery, and accident prevention create low barriers to attack. Colleagues and I 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 could find facility street addresses using publicly available records of the fuel stored onsite for backup power. After demonstrating the feasibility of targeted attacks, we considered several mitigations that could increase the difficulty of a successful attack against access network facilities and infrastructure, and discussed the proposed mitigations with network operators. I plan to continue working to mitigate the risk of targeted attack through further consultation with access network operators.

Illuminating the Opaque 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. My approach to the problem is to build a foundation capable of revealing the networks that cloud traffic traverses and the network facility locations where clouds hand off packets, specific to each combination of cloud datacenter and potential user.

Gathering reliable raw path data from inside cloud datacenters is the cornerstone of comprehensive cloud path analysis, but I found that cloud network routing behavior reduced the accuracy of my previous data science techniques and forced me to adapt existing tools. Conventional traceroute methods accelerate probing with parallelization, and I discovered that the frequent internal route changes inside cloud networks can corrupt many measured paths simultaneously. These corrupted paths create significant noise that is difficult to remove, so I implemented a different probing strategy that rapidly conducts traceroutes with only a single measurement in flight at a time. Through a controlled experiment, I showed that this probing strategy significantly improved the signal-to-noise ratio of the raw data and led to more accurate inferences of the networks along the measured paths.

Long paths over the public Internet can diminish application quality of experience, and I used the higher-fidelity data to design a technique to discover which of the myriad cloud facilities hand off traffic along each

of the measured paths. The geolocation technique I created uses round-trip time measurements to first infer that groups of cloud border router IP addresses reside in the same city, and then leverages interconnection constraints to infer the specific city for each group. 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. Validation on Azure's network indicates that I inferred exit locations with high accuracy.

I found that the interconnection constraints were insufficient to comprehensively geolocate infrastructure outside the cloud. I determined that I could provide better insight into the rest of the path by learning to extract information that network operators embed in their DNS hostnames associated with router IP addresses. These hostnames often include substrings that encode the geographic location and the network that operates the router, but each network uses their own conventions for encoding information. I sought out researchers working on this problem, and together we trained a supervised machine learning model with labels derived from round-trip time measurements and the output of my dissertation techniques. We learned conventions for extracting network operators and geographic locations from hostnames assigned by hundreds of networks, and operators at 13 networks confirmed the accuracy of our learned conventions.

Future Research Directions

The future appears bright for the field of Internet measurement. Along with colleagues, I helped successfully convince the National Science Foundation (NSF) to create a program dedicated to developing new approaches for measuring the Internet, including new data collection methods and data science techniques. 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 benefits of Internet measurement, and are building research groups to better manage their global networks.

The nascent Internet measurement discipline has substantial promise for solving problems in the Internet and provides a fertile ground for the design of transformative new data collection and data science approaches. The techniques I built to extract network interconnections, evaluate access networks, and reveal cloud paths help unblock critically important challenges in Internet research. Among those challenges, I am most excited about three transformative research directions: (1) evaluating the security exposure of Internet paths, (2) assessing the resiliency of critical Internet services, and (3) quantifying broadband performance disparities.

Evaluating Security Exposure of Internet Paths My research program is motivated in part by the long-standing challenge of identifying the countries and networks that Internet traffic traverses. The inability to accurately obtain this information prevents analyses of traffic sovereignty and hinders using the Internet for mission-oriented military communication. After I gave 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—exposing the traffic to espionage, tampering, or disruption. Current tools do not support scientifically assessing the extent of the problem, identifying root causes, and devising actionable solutions.

I plan to collaborate with CIRA to create a generalizable foundation to study traffic sovereignty, with potential implications for international relations and political science. The new insights I bring to the problem are novel techniques for revealing Internet paths, building machine learning models to infer network interconnections and geographic locations, and identifying cloud network infrastructure locations. The foundation I build to study traffic sovereignty will also help support mission-oriented communication over the Internet, and I plan to use the techniques I design to identify Internet paths that only traverse infrastructure in networks and locations acceptable to the traffic source.

Measuring Resiliency of Centralized Services Recent high-profile cloud datacenter outages highlight the risks of consolidating critical services into a handful of networks. I hope to measure the robustness of cloud applications and quantify resiliency successes and failures across different layers, technologies, services, and

applications. The novel contribution I bring to the problem is the foundation I am building to measure cloud application paths. I plan to synthesize probing sources inside cloud datacenters with thousands of probing sources around the world to replicate the process of human users connecting to distributed and replicated cloud applications. The measurements will focus on application reachability and round-trip latency from the devices outside the cloud. My cloud path measurements will help me identify the infrastructure responsible for unresponsive users and applications, as well as correlate latency increases with path changes due to network infrastructure failure; e.g., when application traffic enters or exits the cloud network in different locations or reaches different cloud datacenters.

Revealing Broadband Infrastructure Performance Bottlenecks Despite decades of funding for broadband expansion in the U.S., we cannot properly evaluate the success of these programs or the extent of broadband infrastructure deployment inequities. Existing tools cannot identify bottlenecks in broadband network infrastructure, hampering attempts by policymakers to properly target broadband expansion. I bring to the problem my unique approach for locating network infrastructure failures using reachability measurements to end-user devices. Rather than detect unresponsiveness, I plan to generalize that approach to detect elevated round-trip-times caused by congestion. Using the infrastructure maps I created for the broadband access networks, I will repurpose my existing techniques to identify and locate the source of the congestion, identifying infrastructure bottlenecks where additional resource allocation could provide the greatest performance benefit. I will also combine infrastructure bottleneck locations with socioeconomic status and geography to inform a clear-eyed assessment of Internet infrastructure policy, and create metrics to gauge the success of future broadband initiatives.