Internet security and performance problems often persist invisibly. Operators hide services that later turn out vulnerable, networks sporadically balance load through egregiously-long routing paths, and attackers avoid the network monitors used to capture them. Protecting and improving the Internet's operations requires near-real-time visibility into how networks, and their users, function in practice. Yet, there exists no complete ground truth on how the Internet operates because service configurations, user patterns, and network architectures are diverse, distributed, and ephemeral. There do exist systems that collect data and models that predict operations, but they rely on simplified or unvalidated assumptions about networks, operators, and attackers, restricting their visibility. Consequently, performance troubles and security vulnerabilities exist undetected, allowing them to evolve into catastrophic outages and breaches.

My research builds data collection systems that rigorously account for the Internet's complexities to reveal in near-real time previously-unseen critical operational challenges and threats. I investigate the complex behaviors of networks, attackers, and operators by using quantitative analysis, including non-parametric statistics, to validate and challenge community assumptions—e.g., do applications abide by RFC standards and do attackers take advantage of open source intelligence tools? I build high-performance Internet scanning systems that learn from Internet service deployment patterns to provide the most accurate, global, and up-to-date view of different address spaces and network architectures of the Internet. I then use the data my systems collect to improve performance, defend against attackers, and develop empirical methodologies for future Internet analysis.

The operator, attacker, and network patterns that I identified, along with my systems, have changed how researchers and industry search for performance faults and security threats on the Internet. My approaches—published across top-tier venues (e.g., SIGCOMM, USENIX Security, IMC)—find *billions* of previously-unseen services [1], global performance trends [2], and sophisticated attackers [3]. Governments (e.g., US, UK) and companies (e.g., Hunt Intelligence, Peakhour, Censys) rely on my systems [1, 4, 5] to proactively notify operators about vulnerabilities. Netflix is improving video delivery for over one million satellite users due to the global performance trends surfaced by my system [2]. As the Internet's complexity continues to increase, building data-collecting systems that provide complete and real-time Internet visibility is critical to prevent the Internet's largest performance and security failures from invisibly brewing.

1 Understanding the Internet's Complexity. To build systems that provide visibility into the Internet's performance and security, we must first understand how the Internet operates in practice. I apply an empirical and statistical approach to surface trends and to challenge community assumptions about Internet service deployment patterns, network architectures, and attacker methodologies. My work shows that rigorously modeling complexity is key to quickly finding the most vulnerable and worst-performing parts of the Internet.

Service Deployment Patterns. Internet services follow unexpected deployment patterns that contradict community standards. Prior to 2021, hundreds of Internet studies assumed that IPv4 services live on their assigned ports. My work showed that the vast majority of Internet services do not run on assigned ports [1]. For example, only 3% of HTTP and 6% of TLS services run on their assigned ports 80 and 443, respectively. By scanning for only assigned protocols, existing studies missed roughly 1.9 *billion* (63% of all) services. Worryingly, services on unassigned ports are up to five times more likely to be vulnerable. My work investigating cloud storage found a similar

deployment pattern [6]: widely-used cloud-storage scanning techniques only looked for addresses with short and easy-to-guess names. Vulnerable cloud storage are five times more likely to occupy addresses with high-entropy random strings, causing the security industry to dramatically overestimate the security of cloud storage.

Network Architecture. The operational complexities of networks cause them to behave in unintended and unexpected ways. Recently, we empirically studied the network architecture of Starlink, a low earth orbit satellite network, which uses inter-satellite lasers (ISLs) to route traffic at the speed of light in space. Theoretical models assumed and Starlink marketing insinuated that ISLs could out-compete terrestrial ISPs. In practice, by systematically measuring the latency of LEO satellite users and searching for statistical anomalies, we found that ISLs significantly increase the routing path length (and thus, latency) due to operational satellite constraints [2]. Counter to what theoretical models assumed and network advertisements sold, LEO satellite networks experience an often prohibitive tail latency that jumps ten fold, causing applications such as video conferencing and gaming to be unavailable to users.

Attacker Methodologies. The increased sophistication of both benign and malicious Internet scanning breaks community assumptions about expected traffic distributions and patterns. Using non-parametric statistical tests, we compared attacker behavior across 8 networks and 23 countries and found that attackers that target vulnerable IPv4 services across cloud providers actively avoid passive network telescopes [3]. Consequently, dozens of academic studies that rely on telescopes for attacker visibility miss up to 96% of actively malicious IPs [3]. We also found that attackers mine publicly available Internet-service search engines to find exploitable services. In a separate study, we found that, amongst the noise of benign scanning, malicious scanners are more likely to target cloud storage addresses that belong to bug bounty programs [7].

2 Building Systems for Internet Visibility. Operator deployment, network architectures, and attacker preferences follow complex patterns that traditional data collecting systems are not built to find. To provide visibility into complex ecosystems, my systems leverage concepts from mathematics and machine learning (to predict the location of services and attackers) [4, 6, 8, 9], parallel computing (to build high performance scanners) [5, 9, 4], and computer networking (to manipulate the responsivity of hosts and attackers) [1, 10].

Operator Visibility. Operators host the majority of services on unexpected ports, which hinders visibility into operator behavior [1]. Before my work, no systems scaled to study service deployment across all applications and ports on IPv4. To shrink the discovery of IPv4 services from years to hours, we built GPS, a system that parallelizes probability calculations to efficiently and accurately predict service presence across all ports [4]. To shrink the discovery of unexpected applications by an order of magnitude, we built LZR, a scanner that sends special packets that solicit fingerprintable application layer responses [1]. To continuously provide the most up-to-date view of the global Internet available, we built Kronos [9], a system that operates a continuously learning feedback loop and discovers 94% of IPv4 services in less than 4 hours after they appear—two orders of magnitude faster than prior work. We have also built Stratosphere, a system that uses neural networks to predict the complex names of cloud storage buckets deployment [6], and a system that uses extreme parallelism and a novel caching algorithm to efficiently query domain names [5].

Attacker Visibility. Attackers actively avoid identifiable network monitors, thereby reducing visibility into attacker patterns [3]. Across multiple studies, we built a network of honeypots across

cloud and education networks to attract attacker behavior. For example, we deployed a range of "honeybuckets" on the AWS S3 platform, configured with names, permissions, and content to lure and measure different scanning strategies [7]. Our network is the first to capture the methodology, including which scanning tools, attackers use to find insecure cloud storage.

Network Visibility. Low Earth Orbit (LEO) satellite networks are quickly gaining traction, with over 2 million geographically distributed users to date. Yet, LEO networks remain largely obscure due to the requirement of owning specialized hardware. We built the first system, HitchHiking [2], which measures satellite Internet connections without needing specialized hardware in order to perform measurement studies. HitchHiking builds on the observation that thousands of already-exposed Internet services use LEO Internet, for which the associated routing path reveals satellite network architecture and performance. HitchHiking measures an order of magnitude more LEO-networked paths than state-of-the-art techniques.

3 Improving Performance and Security. The data my systems collect, and trends they surface, help prevent future exploitation and improve the Internet's performance.

Preventing Future Exploitation. Governments and security companies are using my systems to identify and pro-actively notify owners of vulnerable services. Our discovery of widespread service deployment on unassigned ports [1] has fundamentally changed how entities search for vulnerabilities; vulnerability disclosures now prompt vulnerable-device searches across all 65K ports, instead of just the vulnerable protocol's assigned port. My systems have identified vulnerable assets in cloud storage belonging to the department of defense [6] and hundreds of invalid certification authority authorization configurations in the domain name system [5]. Our disclosures have led to the take-down of exposed data, organization internal audits, and DNS registrar re-configurations. My work also provides service and network operator recommendations for how to minimize attracting unsolicited attacker traffic [3].

Improving Video Streaming Performance. The global satellite latency and outage trends that HitchHiking [2] surfaced, along with the inherently challenging nature of sending video over terrestrial networks [11], has prompted Netflix to investigate and improve video streaming for millions of satellite users. Netflix uses HitchHiking's global perspective to isolate which geographic regions are more at risk to subpar video streaming quality of experience. Netflix is on track to completely re-design video streaming for low earth orbit satellite networked users within the next year.

4 Future Work. As the Internet's complexity continues to increase, and attackers become more elusive, building data-collecting systems that provide complete and real-time Internet visibility will become more critical to detect performance and security problems early. I plan to continue building systems that provide visibility into never-before-seen areas of the Internet—both terrestrially and in space—to help us understand where the Internet's most urgent problems lie, how we should fix them, and what is the best way to prevent future performance faults and threats.

Visibility into New Ecosystems. Our visibility of the Internet's ecosystem is shrinking as the Internet expands its address space, adopts new protocols, and implements new privacy measures. Finding and fixing vulnerabilities will become more challenging than ever before. Hundreds of millions of devices are transitioning to use the IPv6 address space and/or rely on name-based virtual hosting; we must build new systems that efficiently pinpoint vulnerable and malicious hosts hidden behind IPv6 addresses and sub-domains. We must gain long-overdue visibility into the increasing adoption of UDP by optimizing systems to detect UDP applications on unassigned ports. We must

develop new solutions that allow us to identify vulnerable and malicious hosts, while respecting new privacy measures (e.g., TLS encrypted SNI, DNS over HTTPS, MAC randomization). In the long term, I would like to shift from creating individually-tailored visibility solutions to a generalizable and systematic framework that can automatically learn and predict patterns across new ecosystems.

Adapting The Internet To Satellite Routing. Low earth orbit satellite networks have quickly become an essential architecture for providing Internet connectivity. However, LEO satellite networks exhibit new performance dynamics (e.g., periodic outages, a long tail of prohibitive latency) that today's Internet applications and protocols are not designed to accommodate. Similar to the onset of high-speed Ethernet, WiFi, and broadband wireless deployment, we must understand the operational complexities of this new network and re-design the network stack accordingly. I want to build the research community's first LEO-specific platform for providing internal visibility into LEO network operations and performance. I would like to develop simulations that are informed by complex operational requirements, in order to explore scenarios that are impossible to test. I want to create algorithmic solutions for improving the detrimental tail performance that LEO networks suffer from, by designing performance enhancing proxies, congestion control algorithms that are aware of LEO-specific network behaviors, and tail-robust applications that make use of buffering. I want to push for and design a standard protocol for data transfer over LEO (e.g., similar to DOCSIS and 5G), to improve researcher visibility into LEO operations for generations to come.

References

† indicates mentee, ∗ indicates co-first authorship

[1] **Liz Izhikevich**, Renata Teixeira, and Zakir Durumeric. LZR: Identifying Unexpected Internet Services. In *30th USENIX Security Symposium*, 2021.

[2] **Liz Izhikevich**, Manda Tran†, Katherine Izhikevich†, Gautam Akiwate, and Zakir Durumeric. Democratizing LEO Satellite Network Measurement. *Under submission to ACM SIGMETRICS*, https://arxiv.org/abs/2306.07469.

[3] **Liz Izhikevich**, Manda Tran†, Michalis Kallitsis, Aurore Fass, and Zakir Durumeric. Cloud Watching: Understanding Attacks Against Cloud-Hosted Services. In *Proceedings of the 23rd ACM Internet Measurement Conference*, 2023.

[4] **Liz Izhikevich**, Renata Teixeira, and Zakir Durumeric. Predicting IPv4 Services Across All Ports. In *Proceedings of the ACM SIGCOMM Conference*, 2022.

[5] **Liz Izhikevich**, Gautam Akiwate, Briana Berger†, Spencer Drakontaidis†, Anna Ascheman†, Paul Pearce, David Adrian, and Zakir Durumeric. ZDNS: A Fast DNS Toolkit for Internet Measurement. In *Proceedings of the 22nd ACM Internet Measurement Conference*, 2022 ⋆**Community Contribution Award**⋆.

[6] Jack Cable*†, Drew Gregory*†, **Liz Izhikevich***, and Zakir Durumeric. Stratosphere: Finding Vulnerable Cloud Storage Buckets. In *Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses*, 2021.

[7] Katherine Izhikevich, Geoffrey Voelker, Stefan Savage, and **Liz Izhikevich**. Using Honeybuckets to Characterize Serverless Storage Scanning in the Wild. *Under submission to Euro S&P*, 2023.

[8] Maya Ziv†, **Liz Izhikevich**, Kimberly Ruth, Katherine Izhikevich†, and Zakir Durumeric. ASdb: A System for Classifying Owners of Autonomous Systems. In *Proceedings of the 21st ACM Internet Measurement Conference*, 2021.

[9] **Liz Izhikevich**, Renata Teixeira, and Zakir Durumeric. Kronos: A System for Adaptively Tracking Internet Service Dynamics. *Under submission to NSDI*, 2023.

[10] Gerry Wan, **Liz Izhikevich**, David Adrian, Katsunari Yoshioka, Ralph Holz, Christian Rossow, and Zakir Durumeric. On the Origin of Scanning: The Impact of Location on Internet-Wide Scans. In *ACM Internet Measurement Conference*, 2020.

[11] Lixiang Ao, **Liz Izhikevich**, Geoffrey M. Voelker, and George Porter. Sprocket: A Serverless Video Processing Framework. In *Proceedings of the Ninth ACM Symposium on Cloud Computing* , 2018.