Project Description: Mid-scale RI-1 (M1-DP): Designing a Global Measurement Infrastructure to Improve Internet Security (GMI3S)

1 Introduction

We propose a Mid-Scale Design Project to design and prototype a distributed but integrated infrastructure to measure the Internet, with the objective of improving Internet infrastructure security.

The Internet's central role in society was demonstrated vividly in 2020. While the Internet has become critical infrastructure permeating all aspects of modern society, its security and trustworthy character are subject to constant threats and attacks. The security of the Internet is a high priority for the security research community, but that community is greatly hindered by a lack of relevant data. Researchers, governments and advocates for society need a more rigorous understanding of the Internet ecosystem, a need made more urgent by the rising influence of adversarial actors. We cannot secure what we do not understand, and we cannot understand what we do not measure. As we both design the future Internet for future generations and operate the current Internet, data is lacking. Through the lens of defense systems analysis, observation (and the infrastructure to support those observations) is the basis of all defense systems [1]. We therefore identify Internet measurement, data curation and making data usable by the research community as critical research infrastructure.

We recognize the need for an infrastructure project to support measurement of the global Internet, similar to how governments support large-scale measurements of the oceans, atmosphere, and various critical infrastructure. But the Internet sits in contrast to other critical systems, such as health care, transportation, agriculture, and commerce, where the government plays a role that complements the role of the private sector – it monitors the state of those systems, and acts as necessary to ensure that they are meeting the needs of society. The first step in this process is gathering data to understand how the system is actually working. Unfortunately, far more than other domains, the scientific enterprise of Internet security is mired in interdisciplinary challenges: complexity and scale of the infrastructure; information-hiding properties of the routing system; security and commercial sensitivities; costs of storing and processing the data; and lack of incentives to gather or share data in the first place, including cost-effective ways to use it operationally.

As a result, today, operators, policy makers and citizens have no consensus view of the Internet to drive decision-making, understand the implications of current or new policies, assess the resilience of the Internet infrastructure in times of crisis, or know if the Internet is being operated in the best interests of society. Governments could gather data directly, but the trans-national character of the Internet raises challenges for government coordination. An accepted approach to data gathering and analysis is to make sure that data is made available to neutral third-parties such as academic researchers, who can independently pursue their efforts, draw their own conclusions, subject these to comparison and peer review, and present their results as advice to governments.

Although we come to this challenge with open eyes, we recognize the scope of the aspiration, and thus propose a substantial 3-year MSR1 Design Project to design a Global Measurement Infrastructure to Improve Internet Security (hereafter, GMI3S).

We do not intend to tackle all Internet security problems. The Internet has a layered structure, which (in its simplest form) is a data transport layer on top of which run a wide range of applications. Our focus is on the Internet as a data transport service, and vulnerabilities specific to that layer: attacks on Internet routing that deflect traffic to bogus destinations (a persistent problem), abuses of the Domain Name System (a widespread, pernicious problem), attacks on the key management system (the Certificate Authorities) that underpin identity and authentication on the

Internet, and spoofing of Internet addresses to disrupt regions of the Internet with untraceable traffic (Denial of Service attacks). Security challenges at these layers seem to get less publicity than attacks on end-points (malware, ransomware, etc.), or design features in applications that lead to risky user experiences. But the challenges at the data transport layer are foundational: they affect the reliable operation of every application that operates over the Internet. Any enterprise or service can have traffic intended for it deflected to a masquerading site that attempts to mimic the legitimate site, steal user credentials, disrupt security, or defraud users (§3). This perhaps surprising fact is the critical driver for this proposal.

Thus, the immediate target for this infrastructure is the research community that measures the Internet and tries to improve its security, the intermediate beneficiary will be the operator community and the service providers of the Internet, and the ultimate beneficiary will be all of society. We recognize that better data alone will not improve the security of the Internet–this proposal is part of a larger community agenda of research and outreach to industry and governments. The proposed infrastructure and its community of users will enable wide engagement of academic groups as well as private-sector security researchers in developing innovative, efficient, and robust capabilities to tackle the challenges of known and emerging Internet vulnerabilities.

Since the Internet is a designed artifact (as opposed to a natural phenomenon of nature), it might seem that one could understand its operational character from analysis of its specifications. This is not so. The Internet is composed of tens of thousands of independent networks and the overall behavior of the Internet is determined by the independent decisions of the operators of those networks. Moreover, in most of the world, the Internet infrastructure is the product of the private sector. Economic considerations that drive the private sector shape the character of the Internet, key aspects of its resilience, security, privacy, and its overall future trajectory. The only way to understand the behavior of the Internet is to measure it.

Consistent with NSF's Blueprint for a National Cyberinfrastructure Ecosystem [2], we view infrastructure holistically, and aim to integrate a range of resources, from sustainable, production data acquisition, to tools for curation, meta-data generation, efficient storage and dissemination, to community services to support accessibility and extensibility of the infrastructure, to domain expertise to enable use of the infrastructure for transformative discoveries.

We organize our proposed work into four tasks. Our first task is to design, prototype, test and evaluate a new highly distributed network measurement platform capable of capturing several types of data relevant to security research, as well as hosting new vetted experiments. This task will require consideration of both dedicated hardware and virtualized software deployments, in a modular architecture that allows hosting sites to opt in to measurements as policy allows.

Our second task includes many facets of data management: meta-data ontologies; standardizing data exchange formats; tools to support data curation and documentation; and techniques for efficient data sharing, discovery, use, and dissemination.

Our third task focuses on community-oriented infrastructure that will enable use of the data for a broad set of cybersecurity research and beyond. This task will tackle issues with sensitive data that raises privacy or corporate concerns. One subtask is to bridge the current gap between the emerging data disclosure control technologies and measurement and security practitioners. We will explore the relevance of computer science advances such as differential privacy and secure multi-party computation, to current and emerging cybersecurity research priorities. We will design a set of legal enablers, e.g., normalization of data-sharing agreements, and socialize these among our partners and the larger community as part of our fourth task, outreach. Task four will include workshops, curriculum development, and STEM/cybersecurity work force training.

To prototype our design, we will work with the community of Research and Educations (R&E) networks, which interconnect campuses and research centers across the globe. The largest R&E

networks in the U.S. and the EU (Internet2 and GEANT), along with ten other academic networks, have agreed to collaborate for testing and evaluation. This is a Design proposal, so many details are as yet unresolved. Reaching agreement on the specifics of the design, informed by prototype deployments, and finding and documenting working solutions is exactly the scope of this project.

2 Intellectual Merit

The *scientific justification* of this infrastructure is enabling a new generation of evidence-based research that can lead to improved Internet infrastructure security. We will design and prototype devices that connect to the Internet to form a platform for distributed measurement. Our proposed platform will collect data directly relevant to the vulnerabilities of these infrastructure layers, at a scale and quality not before available. Past work by the community will help to identify the highest priority data collection objectives. Our goal is to design an infrastructure based on an assessment of the critical infrastructure security challenges, and thus provide a well-structured justification for the priorities for the hardware, software, data and enablers on which we focus.

There is prior work collecting and exploiting such data (§3). This data has informed research related to network infrastructure vulnerabilities, but also supported security research more broadly, including studies of resistance to censorship, human trafficking, child sexual abuse, and Bitcoin fraud. But past efforts have been based on piecemeal measurement, with fragmentary and limited funding, and no ability to construct a long term plan for stable collection, curation, and analysis. Individual researchers devise clever ways to gather data, and with luck find and publish important results describing a moment in time. But individual researchers cannot sustain measurement for decades, or usually even beyond the life of a grant. Our goal for this project is to provide a community-wide cyberinfrastructure that relieves individual researchers of the necessity of developing a transient data collection plan as a part of a single research award, and puts data collection and curation on a sustainable, scalable footing.

The ultimate outcome of our project is a community-vetted and validated design for an innovative infrastructure that will significantly and sustainably advance the nations research capabilities. This cyberinfrastructure will support collection, curation, archiving, and most of all expanded sharing of data needed to answer critical questions about Internet vulnerabilities. In order to enable new discoveries that cut across data sets and domains, our design will leverage commercial network security expertise, apply advanced data architecture and data science techniques, and utilize novel technologies to manage data integrity, availability, and privacy. This proposal targets a long-standing high-priority need in the scientific community that has risen to national importance. Lack of researcher access to data in the U.S. is damaging U.S. competitiveness in a global research environment.

We recognize that data alone will not improve the security of the Internet. Our convergent design approach will cultivate partnerships across academia, industry, and government to establish quantitative metrics for improved national security. Our view is that any successful path forward for Internet security will require a systematic approach to transparency. We believe measurement infrastructure is essential to designing a feasible path toward such approaches [3, 4, 5, 6].

2.1 Community Need

Community workshops have highlighted the need for a new approach such as the one we propose.

1. **Cybersecurity as Big Data Science Interactive Workshop**, April 2021, where participants highlighted CAIDA's existing traffic data as critical [7].

2. At NITRD's **Huge Data Workshop** [8] participants observed that huge data problems often arise from continuously generated data of massive volume, such as network traffic.
3. NITRD's **Federal Cybersecurity R&D Strategic Plan** emphasized the importance of evidence-based evaluations and measurements in cybersecurity research, and recommended that the Federal Government prioritize basic and long-term cybersecurity research, including the development of sound scientific foundations and formal, reproducible, and quantifiable methods for assessing the effectiveness and efficiency of cybersecurity solutions.
4. CAIDA's **Active Internet Measurement Systems (AIMS)** workshops [9] have for the last 12 years brought researchers, operators, and government stakeholders together to discuss existing Internet measurement system challenges. A perennial topic at these workshops has been the challenge of supporting longitudinal data [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20].
5. CAIDA's **Workshop on Internet Economics (WIE)** workshops [21] have for the last 11 years to discuss how to overcome data access barriers for economic and policy researchers seeking to study the Internet [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. This series has also consistently observed the decreasing visibility of the infrastructure and the threat it poses to the U.S., not only in a cybersecurity context.
6. The 2017 **EU/US Future Networks** workshop [33] called for building, deploying, and maintaining measurement infrastructures across the Atlantic, to support joint research on defining standard metrics and shared measurement methods for regulating the Internet market, and comparative studies of distributed data sets across regions.
7. In 2021, the NSF-sponsored two **Workshop on Overcoming Barriers to Internet Research (WOMBIR)** workshops [34], where participants heavily emphasized the need for dedicated infrastructure to support longitudinal studies of the Internet.

2.2 Research Community Benefits

Our proposed design targets NSF's recently articulated cyberinfrastructure goals for pilot projects [2], namely: to improve the process of accessing, integrating and transforming data to knowledge and discovery; to enable new usage modes to address multi-disciplinary and cross-domain scientific objectives; to address emerging community-scale scientific data challenges such as real-time, streaming data, data discovery and delivery, privacy sensitivities, data fusion, integration and interoperability; enhancing data accessibility and utility. Beyond the security threats that are our primary science drivers, our data collection will provide valuable information to the data-intensive field of Internet cartography, a sub-discipline of Internet infrastructure research, including mapping router-level topology [35, 36] cable provider performance [37], submarine cables [38, 39], to measuring and analyzing national AS choke points [40, 41, 42], to understanding the mobile Internet topology [43, 44]. Our project also prioritizes technical and operational cybersecurity work force training; data-intensive methods for assessing security, stability, and resilience of network infrastructure; and enabling scientific and engineering advances to navigate other current and future harms to Internet infrastructure.

Machine Learning and AI for Networks The infrastructure that fuels the current wave of AI/ML—the Internet itself—is one of the last to benefit from those advances. The obstacle is neither conceptual nor computational complexity, but access to data that feeds the algorithms. Application of data-driven ML techniques to Internet infrastructure research brings many challenges: each network is unique, dynamic, typically not instrumented for scientific measurement much less manual labeling of data, characterized by anomalies and misbehaviors that complicate the creation of training data sets, and usually proprietary. ML learning paradigms that do not require labeled

training data (unsupervised learning) or at least not perfect data (semi-supervised, reinforcement learning) have received less attention. A 2018 survey of ML for networking found that the vast majority of research use either synthetic data or data that generally does not reflect realistic environments [45]. To quote their conclusion, "a combined effort from both academia and industry is needed, to create public repositories of data traces annotated with ground truth from various real networks." Other ML, networking, and security researchers have echoed this call [46, 47, 48]. Task 2 focused on design components (standard formats and metadata (§5.1,§5.2) that focus on creating data sets that are sufficiently annotated so that ML researchers can use them to train ML models.

3 The Critical Security Flaws in the Internet

In this section, we describe in more detail the security threats that constitute the primary science driver for our infrastructure, and the state of community research in these areas. We offer evidence that progress is possible with the right combination of improved infrastructure and revitalized research, and we argue that now is the critical time to undertake this effort.

The function of the Internet transport layer is to deliver packets of data [49]. The operation of the Internet transport layer depends on several critical system elements.

- *Internet (IP) addresses*: every element communicating across the packet layer, i.e., using IP protocols, including end-points and routers, is assigned one or more addresses, so that packets can be delivered to it.
- The *global routing protocol* (the Border Gateway Protocol or BGP) [50] propagates topology and routing policy information across 70K+ independent networks called *autonomous systems*. This information enables routers to correctly forward packets to their destinations.
- The Domain Name System or DNS translates human-meaningful names (like www.nsf.gov) into IP addresses to which routers forward packets. If this system is working in a trustworthy manner, the user will obtain the correct IP address for the intended higher-layer service, e.g., web site, and will not be misled into going to unintended or malicious locations.
- The Certificate Authority system manages and distributes to users encryption keys used for transport connections so that users can confirm the identity of the party with which they are communicating. If this system is working correctly, the user receives a confirmation that the service at the end point receiving the packet is the service the user intended to reach.

To simplify, if these systems are working correctly, the Internet as a packet forwarding system – its "plumbing" – is working correctly. Unfortunately, all of these systems suffer from known vulnerabilities [51, 52, 53, 54, 55], which attackers regularly exploit [56, 57, 58], despite decades of attempts to remediate them.

3.1 Internet Addressing System (IP)

The network layer of the Internet architecture is most fundamentally defined by Internet Protocol (IP) addresses. Routers use destination IP addresses in the headers of packets to choose the next hop to forward a packet toward its intended destination. A well-known vulnerability embedded in this layer is the ability to *spoof source IP addresses*. To reach its destination, a packet must obviously have the destination's IP address in its header. Similarly, if the destination is to return a packet to the original source for two-way communication, the source address listed in the first packet must correspond to the actual source of the packet. But if a malicious source sends a packet to a destination using a fake source address, e.g., one belonging to a third endpoint, the receiver of the packet will reply to that third endpoint's address rather than to to the original sender. In

fact the receiver cannot respond to the original sender since it does not know the actual source; it trusts the authenticity of the source address field in the header. Malicious actors have exploited this vulnerability to mount a variety of attacks, e.g., volumetric denial-of-service [58] (DoS), resource exhaustion, cache poisoning, and impersonation [59].

In the early days of the Internet, the designers appreciated this risk in principle but the idea of an attacker subverting perhaps hundreds of thousands of end points for malicious purposes seemed remote. Today, attacks that involve hundreds of thousands of machines [60, 61], with terabits of malicious traffic [58], are regular events on the Internet [57]. The problem is exacerbated with the exhaustion of IPv4 space and subsequent emergence of grey market, making it more challenging to determine address ownership [62].

3.2 Internet Routing System (BGP)

There are about 70K Autonomous Systems that make up the Internet today. Each AS may own a set of IP addresses, and every AS in the Internet must know how to forward packets to these addresses. The Border Gateway Protocol, or BGP, is the mechanism that ASes use to propagate this knowledge across the network topology. Addresses are organized into *address blocks* of various sizes, identified by the *prefix* (the first part) of the addresses in the block. Each AS uses BGP to announce to its directly connected neighbor ASes the prefixes that it hosts. The receiving ASes pass this announcement on to their neighbors, and so on, until (in principle) it reaches all parts of the Internet. As each AS passes an announcement along, it adds its own AS number to the announcement, so the form of the announcement is a series of AS numbers that describe the path (at the AS level) back to the AS owning the associated address block.

The critical security flaw with BGP is well-known: a rogue Autonomous System can announce a falsehood into the global routing system, i.e., a false announcement that it hosts or is the path to a block of addresses that it does not have the authority to announce. Traffic addressed to that block may travel to the rogue AS, which can drop, inspect, or manipulate that traffic. The simplest form of the resulting harm is that traffic goes to the wrong part of the Internet, and is then (in the best case) discarded. This outcome leads to a loss of availability between the parties intending to communicate. A more pernicious kind of harm is that a rogue endpoint can mimic the behavior of the intended endpoint, and carry out an exchange that seems to the victim to be with a legitimate party. This attack can lead to theft of information such as user credentials, which the malicious actor can then exploit. Hijacks can also lead to the download of malicious software, or malware, onto the victim's computer. Another possible harm is that the malicious actor may launch some abusive traffic from addresses in that block, damaging the reputation of the owner of the block.[1] A recent study described how BGP routing attacks can de-anonymize Tor participants disrupt the cryptographic underpinnings of the Internet, and destabilize the BitCoin blockchain [63].

Many academic studies have used data-driven methods to understand and propose mitigations to the insecurity of BGP. They have relied either on raw data from NSRC RouteViews (project partner) and RIPE RIS (collaborator), or on data feeds derived from these two data sources [64, 65, 66, 54, 67, 68, 69, 70]. As of April 2021, RouteViews has 33 hardware collection devices that gather routing data from 318 Autonomous Systems by maintaining over 800 peering sessions. (Some ASes peer with multiple collectors.) Insufficient coverage has always been an obstacle to research with this data, since density of vantage points is critical to inference accuracy. To mitigate coverage gaps, other researchers have deployed active probing from vantage points across the Internet

[1] This attack may seem to be an abuse of the addressing system, but it is the routing system that allows one user to appropriate another user's addresses. Spammers will hijack a small block of addresses, send a large volume of spam as the impersonating network, and then withdraw the hijack.

to detect routing anomalies. iSPY [71] depended on having a probing point within each AS that wanted to detect if its routes are being hijacked. Many Internet measurement studies used PlanetLab [72], but PlanetLab has been decommissioned, leaving the research community without an obvious strategy to experiment with active probing. For example, two early routing anomaly detection projects [73, 74] depended on both PlanetLab and RouteViews data feed. Other projects required an active probing platform [75]. The creators were not able to keep these in operation.

Recently, MIT and CAIDA [76] developed a scheme to identify malicious routing announcements based on their intrinsic characteristics. Working with five years of RouteViews data curated by CAIDA, they demonstrated that there are Autonomous Systems that persist as malicious players in the Internet for years, issuing malicious routing announcements and deflecting ("hijacking") traffic away from its intended destination. They identified about 400 of the 70K active ASes as highly likely serial hijackers, and another 400 that are probable hijackers. Similarly, third-party related work used RouteViews data with CAIDA routing annotations [77] to identify bulletproof hosting ASes as part of an AS reputation system [78].

After decades of failure to solve the routing security problem by the design of new protocols and standardized mechanisms [79], there is growing consensus that the path to improved security will instead involve commitments from network operators and other critical actors to undertake enhanced operational practices that restrict the ability of malicious actors to disrupt normal activities. The role of measurement is critical at three points in this process. First, measurement is required to study which practices will yield improved security in practice. Second, measurement is important to demonstrating that the benefits justify the investment by different actors in the ecosystem. Finally, measurement is essential to monitoring compliance with these commitments.

As an example of enhanced practices, a group of network operators, facilitated by the Internet Society, have defined a set of operational practices that can prevent several types of addressing and routing abuse (§3.1 and §3.2). This code of conduct, launched in 2014, is called the *Mutually Agreed Norms for Routing Security (MANRS)* initiative [80]. To ensure progress, members who commit to the MANRS practices must actually undertake them. Currently, the Internet Society has no infrastructure to measure compliance. For example, CAIDA has provided data for their evaluation of the MANRS source address validation filtering practice. However, CAIDA's existing tools cannot scale to the level necessary to provide compliance checking if and as MANRS is more widely adapted. Our data collection can lay the foundation for enhancing the MANRS practices over time, and as well serve to verify compliance with other aspects of those practices. The Internet Society, as the steward of the MANRS initiative, will work with us to identify the highest priorities for data collection, and integrate our data into their practices (see LoC).

3.3 Internet Naming System (DNS)

The DNS translates a name of the form `www.example.com` into an Internet destination address to use in the packet header to forward the packet. Structurally, the DNS is a hierarchical, distributed database with built-in redundancy. Assignment of responsibility for domains occurs through a process of *delegation*, in which an entity at a higher level in the hierarchy assigns responsibility for a subset of names to another party. The hierarchy starts at the *root* of the DNS, which delegates top-level domains (TLDs) such as `.com`, `.net`, `.nl`, etc. to *TLD registries*. These TLD registries in turn delegate to second-level domains, e.g., ucsd.edu, which may further delegate parts of the name space, e.g., cs.ucsd.edu. Administration of these delegations can be a complex task involving many stakeholders, most obviously *registries*, *registrars* and *registrants*. A DNS *registry* administers a TLD. The *registrar* provides an interface between registrant and registry, managing purchase, renewal, changes, expiration, and billing. The *registrant* is the customer that registers a

domain. The organization with overall responsibility for the stewardship of the DNS namespace is the Internet Corporation for Assigned Names and Numbers (ICANN).

Today, harms that leverage the DNS protocols and supply chain represent some of the most pernicious security threats on the Internet, creating daunting policy challenges [81]. Malicious actors can subvert existing names or register their own names by penetrating databases operated by either registries or registrars, and then use those names for malicious purposes. By penetrating a registry or registrar database, one can add invalid registrant information, or change the binding from a name to an address. Lack of oversight of the competitive for-profit DNS supply chain contributes to these security risks. But the complexity of the DNS also leads to misconfiguration of the name resolution mechanisms by owners of domain names, which can allow malicious actors to take control of them. Finally, and most challenging, is the registration of domain names intended for malicious use such as phishing or malware delivery. Phishing attacks, or hijacking DNS resolution transactions to provide false answers, can scam users out of money and identity. Every month the Internet Corporation for Assigned Names and Numbers (ICANN) reports the number of active domain names associated with abusive practices.[2] These range from a low of 572K in July 2020 to a high of 926K in October 2020.

Some registrars support operational practices that seem tailored to the needs of malicious actors, such as automatic registration of bulk, meaningless domain names, the creation on demand of "look-alike" or "impersonation" names, or lax attention to capturing the identity of the registrant. Often the DNS is only one component of malicious activities, and stakeholders disagree on whether the DNS is a suitable or effective system through which to combat them. Evidence would suggest that the many private sector security companies existing today are not demonstrating sufficient progress on security.

Researchers have spent years studying different aspects of DNS security vulnerabilities and abuse, e.g., including dependencies that threaten resilience [82], configuration vulnerabilities [83, 84, 85, 56]. Multiple research groups have found that a large fraction of malicious activity can be attributed to a few malicious actors [86, 87, 88]. These discoveries shed doubt of the blacklisting approach (maintaining vast lists of malicious domain names) as a cost-effective means to achieve security [87]. Other researchers have found the blacklists to be inconsistent in content and format [89], with each other and with other data [90], suggesting the entire security ecosystem needs firmer scientific footing.

But as with BGP, the studies are consistently just one-shot, the current infrastructure does not allow longitudinal studies, stunting progress in the field. The Domain Name system is more complex than BGP, and the path toward better security is less clear. There are more layers of operation, more players in the ecosystem, more tensions among them, and fundamentally more things can go wrong. For BGP, the Internet Society, through MANRS, has already taken the initiative toward a set of operational practices that would reduce the attack surface for abusers of routing system vulnerabilities. It is not clear who could play the role of incentivizing operational norms for the DNS that would enhance security.

The EU's new privacy regulation (GDPR) has further challenged DNS research, as most registrant data that used to be public and essential for cybersecurity research [91] has now been removed from public view in an over-conservative compliance with the regulation. We must address issues of this sort as a part of our overall design process.

[2]This data is reported in the monthly DAAR reports, which can be found at https://www.icann.org/octo-ssr/daar.

3.4 Internet Certificate Authority (CA) System

The Certificate Authority (CA) system plays a critical role in Internet security. When operating correctly, it provides a means for a user (typically via a web browser) to verify that a connection is to the intended destination–e.g., the correct banking site–rather than a rogue copy of the site. However, the CA system itself is vulnerable to attack and manipulation. Some certificate authorities may issue misleading certificates providing the wrong public key (the verification credential) to a user. The assumption behind the design of the current CA system was that all CA authorities would be trustworthy, even in a competitive for-profit environment with no regulatory oversight. Not surprisingly, this has proven false in practice. A variety of recent work has demonstrated the "race to the bottom" nature of private sector provisioning of TLS security [92, 93, 94, 95].

If an attacker can cause the issuance of a invalid certificate, whether by penetrating a CA and subverting it, paying a untrustworthy CA to issue such a certificate, or simply (and in particular for state-level attackers) working with a CA that acts as an agent of the state in issuing false certificates, an attacker can pretend to be an end point that it is not, even if the victim end point uses encryption and authentication to attempt to verify the identity of the other end. This attack complements DNS or BGP hijacks that bring traffic to that rogue end point, which then emulates the expected end point, even to the point of cryptographically identifying as valid.

The Certificate Transparency (CT) project[3] provides a critical source of data about active certificates. Our current view is that our platform can best improve CA security by supporting active probing to complement what is recorded in the CT logs, and to merge information gathered about BGP and DNS together with data about changes in the certificate hierarchy.

4 Task 1: Design, Test, and Evaluate Data Acquisition Node

A core component of our cyberinfrastructure is a widely distributed platform for active and passive network measurement. Conceptually, each node in this platform would be a physical hardware device that would be installed in the data center of a network provider, and connected to their network. In our prototype phase, we contemplate deploying about 20 such monitors over 3 years. To scale our design to over 1000 collection points, we will also design a software-only containerized version, which can be installed on a management platform provided by the operator.

4.1 Design Data Acquisition Component

We propose to design a integrated distributed platform of measurement node that can perform a range of data collection tasks, as well as support new vetted experiments by the research community. This design process will rely on recently developed taxonomies of harms [96, 97] as a basis to identify potential datasets that can shed light on each one, and a concrete proposal for how such datasets might be gathered. Initial data sets we anticipate starting with:

- **Routing data.** A detailed view of Internet routing would ideally include collection of routing information from all 70K+ Autonomous System on the Internet, but diffuse ownership of Internet networks makes this aspiration unrealistic. We aim for a design that can accommodate O(1000) collectors each peering with, and thus gathering views from multiple other networks. The design exercise must include how to optimize vantage point placement to maximize visibility of topology. Collection at this scale will require an entire re-engineering

[3] https://certificate.transparency.dev/

of Route Views, such as employing a data streaming infrastructure [98]. Increasing the number of collectors, in parallel with Internet growth, may require that we handle 100x current rates of incoming routing data.

One challenge of detecting routing hijacks and other anomalies is to determine ground truth, i.e., the actual intentions of the owner of the address block. A practice gaining traction in some parts of the world is to register legitimate uses (i.e., who can announce) of addresses in a registry, using a Route Origin Authorization or ROA [99]. If a network deploys our proposed measurement node, the node will detect (in real time) if the network has misconfigured BGP or ROAs, and are announcing routes that are inconsistent with those ROAs. The nodes will also identify BGP announcements arriving at the network from the Internet that are invalidated by a ROA, thus revealing the degree to which invalid routes are being detected and dropped across the Internet. For ASes that have not yet chosen to register ROAs for their address blocks, our platform could host a system such as DISCO [100], which uses the BGP announcements made by the AS as ground truth, and detects announcements originating elsewhere in the Internet that are inconsistent with the announcements from the owner of the addresses.

- **Unsolicited traffic.**, e.g., scans and attacks that arrive at a device. CAIDA currently operates a collector for unsolicited "background radiation" traffic, called a *network telescope*. This traffic is collected by announcing a set of Internet addresses that are not otherwise used (so they receive no legitimate traffic) and then capturing traffic destined to these addresses. The CAIDA telescope uses a block of addresses that are fixed and thus known to adversaries. Research has shown that different address space observes different spectrum of traffic [101]. Given the exhaustion of IPv4 address space with which to create such instrumentation, the only realistic way to sustain such data collection is to design a distributed telescope capability that captures data across many parts of active Internet address space (analogous to the Very Long Baseline Array [102]). Such an idea involves many design challenges, most notably how to reliably discern unsolicited traffic, but success here would lead to a transformative instrument for security research.
- **Compliance verification.** Our design will include the ability to perform active measurement to verify compliance with MANRS requirements [103]. The design will also consider data from Internet Routing Registries (IRR) and Resource Public Key Infrastructure (RPKI) to validate route announcements from customers.
- **Topology mapping.** Our design will include topology mapping capabilities (using traceroute) to map how the measurement node reaches the rest of the Internet. One can correlate traceroute data with routing and DNS data to detect, geolocate, and analyze routing failures, anomalous configurations, network resilience, and mis-behaving infrastructure. Similar to unsolicited traffic telescopes, research has shown that different active measurement vantage points observe different properties [104], so a distributed platform is valuable.
- **Other measurements.** Other security-relevant measurements we will consider in the design process include DNS lookups, SSL Certificate, and performance measurements [105]. We intend for our design to be able to accommodate other sources of data where the point acquisition does not need to be colocated with a given vantage points.

Extending use of platform to researchers Our design needs to accommodate new measurement projects by vetted researchers. In the past, the PlanetLab infrastructure hosted important network measurement projects, but that system is no longer operational, had resource constraints that limited its utility for measurement experiments, and did not vet experiments.

Overcoming Barriers and Incentivizing Deployment. Deployment of our measurement nodes will require the cooperation of operators. Barriers to deployment include risk of misuse of the data, and costs of deployment and operation. To reduce the first risk, we will select measurements that are not likely to reveal sensitive data, and make it possible for the hosting operator to limit what tests are performed. To provide additional incentive, we will design measurements to provide valuable information back to the hosting AS. Collection at this scale and density will allow us, and the hosting sites themselves, to do a much better job of detecting routing anomalies and attacks [106], failures in compliance with operational practices [59], and other operational anomalies and security threads. We will work with operators to design cost-effective measurements that balance benefits against risk and costs. If we can reduce the costs sufficiently, it may be possible to make deployment of a measurement node a required operational practice, perhaps as part of an enhanced MANRS (§3.2).

4.2 Design and Prototype Virtualization Capabilities

We have described the measurement platform as a set of distributed stand-alone hardware devices. In many cases this will be the preferred embodiment, and one result of our design will be specific hardware options. However, we will also leverage a current trend in network operation called Network Function Virtualization or NFV, which allows network operators to install general purpose computers that host *virtual machines*. Network operators use virtual machines to deploy new services without having to install new hardware. Route Views, RIPE, and CAIDA have demonstrated the use of virtualization for their measurement platforms, expanding deployment options. CAIDA is collaborating with UIUC (LOC) on the PacketLab [107] software project, which offers another virtualization option, albeit with limited measurement capabilities. We will consider these options as part of a modular software architecture.

4.3 Test and Evaluate Prototype Deployment in R&E Networks

As part of our design phase, we need a subset of the Internet where we can work with operators, deploy and evaluate prototypes, and develop practices for community access to our data. The research and education (R&E) networks are ideal candidates for early engagement. The R&E networks encompass campus networks as well as regional, national and trans-national networks focused on the needs of the R&E community. These needs include high performance to support processing of large science data sets, privacy concerns around students, protection of unreleased research results, stability of operation, and so on.

Working with R&E networks has several advantages. First, R&E networks are in general not built by profit-seeking entities, reducing some concerns about adverse consequences of proprietary data release. (But concerns still exist – see §6.3). Second, by targeting this community, our initial conclusions can be of direct benefit to the users of those R&E networks, which support the larger NSF-funded practice of science. Finally, and most importantly, that community has a history of working with the network research community, and an understanding of the importance of network research in furthering their own mission.

The primary national research networks in Europe and the U.S., (GEANT [108] and Internet2 [109], respectively) have agreed to support this prototyping and testing (see LOCs). Other initial collaborators are Great Plains Network, CENIC (California's R&E network), the Pacific Research Platform, UC San Diego, UC Santa Cruz, U. Hawaii, UIUC, U Oregon, and SIDN (in Holland). We will collaborate with the EPOC and NetSage projects as channels to the R&E community networks, and to leverage the PerfSONAR framework where feasible.

4.4 Assessment of Deployment in R&E Networks

Our final outcome for this task will be a comprehensive assessment report that includes the specifications used for testing and evaluation, and the experiences deploying these nodes across R&E networks. As a concrete milestone, we will use the test infrastructure to map R&E cyberinfrastructure with unprecedented detail, i.e., the networks of the R&E community in the U.S. and globally (§4.3). We will use this report to gain wider community exposure and commitment to participate in an Implementation Project proposal.

5 Task 2: Design Infrastructure for Data Management

The GMI3S cannot achieve its objectives if the data and tools it hosts are too difficult to use, both for researchers engaged in the scientific enterprise, and for training the next generation of cyber experts. Anyone who has worked with datasets provided by others knows that there is a learning curve to understand what a given data set represents, the range of validity for various queries, and even how to access and process the data to extract sound insights. Researchers who misunderstand how to interpret data risk drawing invalid conclusions, an amplified risk for new or naive users of the data.

Our second design task thus covers many facets of data management that can reduce the learning curve to exploit the data, and helping researchers – especially students and early career academics – make effective, efficient and valid use of the data. We will design meta-data ontologies; standardize data exchange formats; design and prototype tools to support data curation; and write documentation that includes guidance on valid usage and caveats; and techniques for efficient data sharing and dissemination.

5.1 Design Standard Data Representation and Meta-Data Formats

We will define uniform representations for data that may be heterogeneous in its raw forms. Some traffic captures will come from our own monitors. We hope the legal enablers enable some data to come from industry. As part of the intake of these data feeds, we must establish a uniform representation and associated meta-data so that researchers have a uniform access method to the data. CAIDA is currently collaboration with RIPE and RouteViews on standardizing content and format of BGP data [110]. We also need to consider meta-data to help interpret the gathered data, e.g., geolocation or organizational ownership data, DNS data, data from other research projects, or from private network operators (§6). We will adapt existing standards [111] for documentation of meta-data, formats, and methods of access. Collaborator Joel Sommers will bring his expertise automating measurement metadata collection and analysis to this task [112].

5.2 Design Tools for Curation and Documentation

These standard representations will create increased burdens on data curators. In order to annotate data at scale, we will design systems for automated curation and documentation. We will design new data and metadata APIs and software development libraries that help synthesize and map low-level structural information about the Internet with high-level phenomena investigated by domain researchers. We envision a layered set of APIs (lower level APIs that provide direct access to specific types of data knit together by a higher level API that supports queries that federate and synthesize the data across lower level APIs).

5.3 Design Data Discovery Architecture

To target all four FAIR principles (findable, accessible, interoperable, reusable) [113], we plan to design a rich context framework for the data. We will design and prototype an open source catalog that addresses the challenge of finding, for a given Internet research data set, who has worked with the data before, what methods and code they used, and what results they produced [114]. We will build on our recent attempt to catalog CAIDA's data sets in this manner [115], including snippets to teach users how to process data for common tasks. To evaluate the catalog design, we will adapt this software as needed to index the thousands of Internet infrastructure data sets CAIDA has gathered and/or curated over the last 20 years, including links to all publications that have used each data set, code those authors are willing to share, and associated metadata. In a future Implementation Phase of the GMI3S, we envision this platform indexing a broad and heterogeneous set of data, including proprietary data sets shared by commercial firms, the terms under which they shared it, and API access.

5.4 Design Efficient Dissemination Approaches

Our design process will consider how to leverage existing work on efficient distribution of scientific data, including the use of multicast to broadcast stream of research traffic to R&E networks [116], the Open Storage Network [117], and the Pacific Research Platform (see LOCs). The Unidata LDM-7: a Hybrid Multicast/unicast System for Highly Efficient and Reliable Real-Time Data Distribution, and just completed a prototype deployment on Internet2 [116]. The Open Storage Network includes seven 1.3PB storage nodes at Johns Hopkins, SDSC, NCSA, MGHPCC, RENCI. Allocations are also available via XSEDE. The OSN will provides a common interface to accessing the Ceph object store via an S3 API. The Pacfiic Research Platform is a partnership of more than 50 institutions to create a seamless research platform that encourages collaboration on a broad range of data-intensive fields and projects. Their goal is to achieve transparent and rapid data access among collaborating scientists at multiple institutions that extends the university campus Science DMZ model to a global scale. We will leverage expertise from these projects in our design process.

6 Task 3: Design Community Infrastructure for Broad Usability

Not all relevant data is amenable to collection from our measurement node. In this task we tackle the design challenge of accommodating data from external, including commercial, sources. But capturing Internet data raises important issues around privacy of the users whose data is being captured. In particular, according to the European General Data Protection Regulation (GDPR), Internet addresses are PII. At the same time, IP addresses are critical as elements of data records that allow data to be joined across different data sets. Anonymizing the IP addresses prevents doing this sort of synthesis, and greatly impairs our ability to draw conclusions from the larger picture that require multiple data sets. We will take a phased approach to this task. First we will design technical interfaces to accommodate new data sources 6.1. In parallel we will undertake efforts to explore both technology (§6.2) and policy (§6.3) vehicles to support sharing of sensitive data, and design a hybrid framework that leverages both approaches. Finally, we will use case studies to evaluate and refine our design.

6.1 Design approach to integrate additional data sources

We will begin this task using data we control ourselves, before we tackle privacy issues. We will design interfaces to allow GMI3S users to acquire data through CAIDA's Periscope interface to public looking glass infrastructure [118] to extend visibility for topology mapping. The design objective will be the ability to capture data from multiple data sources using accessible, consistent interfaces. We will also consider other sources of infrastructure data in this phase of the design: OpenIntel active measurements [119]; archived daily TLD DNS zone files [120]; and passive capture of DNS queries from recursive resolvers, i.e, that do not reveal and personal information about a user [121]. We will evaluate our design by prototyping it using the OpenIntel data.

6.2 Design Disclosure Control Approaches: Software

There are known methods to support work with PII, ranging from careful research protocols and codes of conduct (§6.3) to highly technical options such as differential privacy, generative adversarial networks [122, 123], secure multi-party computation [124, 125], database anonymization [126], and others [127]. Unfortunately, these privacy-preservation techniques have proved too steep a learning curve for network and security researchers, leaving a daunting gap between the two fields. As part of this design project we propose to narrow this gap, by dedicating at least one workshop a year to bringing together experts in both fields to find common ground. In concrete terms, we must discover the range of cybersecurity questions that various techniques can support on various data sets. This will require creating a taxonomy of data, including proprietary data, and understanding in more depth the concerns that arise about sharing, so that we can design repeatable practices to enable legitimate research access to various data types.

6.3 Design Disclosure Control Approaches: Policy Tools

There will inevitably be some questions that cannot be investigated with existing privacy-preserving frameworks. Fortunately, there are well-understood practices, used in this and other sectors, to allow access to data by qualified independent scholars in a responsible manner (Table 1), and in a way that allows replication or original research that builds on previous work. In this design phase, we will identify lessons learned from previous data-sharing efforts [128, 129]. Notably, the Menlo Report [130, 131] proposed a summary of principles to guide the identification and resolution of ethical issues in information technology research, and a companion report that applies these principles to real and synthetic case studies. We will also investigate how Europe and other parts of the world are approaching this same issue.

In this design phase, we will pursue agreements with commercial data providers on practices that will not expose them to liability for privacy violations, and that embody the aspiration that what the research community learns from shared data can be valuable to the firm that shares it. We must also consider that with increasing governmental involvement will come increasing pressure to share data. Working out concepts and approaches to sharing can give both industry and academics more of a voice in the methods and expectations around sharing. Attorney Aaron Burstein (LOC), who wrote the most well-known legal scholarship on amending the law to facilitate cybersecurity research [132], has agreed to collaborate with us to design data sharing agreements that balance these concerns.

- Data is made available in curated repositories, or otherwise provided in ways that allows adequate access for legitimate scientific research
- Access requires registration with data source and legitimate research need
- Standard anonymization methods are used where needed
- Recipients agree to not repost corpus
- Recipients agree that they will not deanonymize data
- Recipients can publish analysis and data examples necessary to review research
- Recipients agree to use accepted protocols when revealing sensitive data, such as security vulnerabilities or data on human subjects
- Recipients agree to cite the repository and provide publications back to repository
- Repository can curate enriched products developed by researchers

Table 1: *Codes of conduct have been developed, in the context of Internet research and in other fields, that enable responsible sharing of data in ways that protect stakeholders while allowing research. To keep science and engineering communities competitive, governments will need to encourage and incentivize sharing data under such standard usage agreements. Funding agencies, journals, conferences and professional societies should also incentivize research conducted under these conditions.*

6.4 Demonstrate Extensibility of Policy Framework with Case Studies

A globally distributed measurement platform is a centerpiece of our proposed cyberinfrastructure, but we will try to accommodate other data sources in our infrastructure design. Traffic data allows researchers to look for communication patterns typical of botnets or DDoS attacks, communication with suspect end-points on the Internet, detection of IoT devices [133], and overall trends in network usage. However, packet-level data includes IP addresses, which are Personally Identifying Information (PII) and can reveal problematic business relationships. We will work with two commercial providers (Kentik and Farsight, see LoCs), both of which gather data from a large number of ISPs, to develop a policy framework to make this data available in some form to the research community. The combination of BGP data from our monitors, traffic data from Kentik, passive DNS data, and our various DNS data sets will provide a strong foundation for researchers interested in improving Internet security.

7 Task 4: Infrastructure for Outreach

In Task 4 we will organize and maintain community resources that guide our design: bi-annual workshops and a virtual environment to sustain design conversations between workshops, and curriculum/work force training material.

7.1 Workshops and Collaboration Environment

We will host two workshops each year, which will provide opportunities to develop consensus around the priority of different areas, the data that can underpin research in those areas, and how to best make that data available for research. The workshops will facilitate conversations among academic, private, and public sector actors. These conversations will also be interdisciplinary, to cross-fertilize expertise in network engineering and operations, cybersecurity, cloud computing support for sensitive data, legal expertise in data sharing and ethical impact assessments, and sound policy implications of research results.

To continue design activities in between workshops, we will host a virtual collaboration environment for multi-threaded discussions related to the project. These workshops and virtual environment will both serve to bring together participants from academia, industry, and governments, including a large community of researchers that want to use the resulting data for research. As a basis for these discussion, We will use and evolve our taxonomy of Internet-related harms that map to research priorities and data sets [96].

7.2 Design and Prototype Modules to Scale STEM Work force Development

To scale methods for training students and researchers how to use Internet measurement data, we will develop a Network Infrastructure Data Science course, including modules on responsible use of the data we include in the GMI3S design. We will create a video archive of useful snippets of instruction to add to the project web site. We will also promote the use of the datasets and analytics in cybersecurity curricula.

References

[1] W. P. Delaney, *Perspectives on Defense Systems Analysis*. MIT Press, 2015.

[2] "Transforming Science Through Cyberinfrastructure," Feb. 2019. Draft of NSF's Blueprint for a National Cyberinfrastructure Ecosystem.

[3] David D. Clark and kc claffy, "Trust Zones: A Path to a More Secure Internet Infrastructure," *Journal on Internet Policy*, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3746071.

[4] Hesselman, Cristian and Grosso, Paola and Holz, Ralph and Kuipers, Fernando and Xue, Janet Hui and Jonker, Mattijs and de Ruiter, Joeri and Sperotto, Anna and van Rijswijk-Deij, Roland and Moura, Giovane C. M. and Pras, Aiko and de Laat, Cees, "A Responsible Internet to Increase Trust in the Digital World," *Journal of Network and Systems Management*, 2020. https://doi.org/10.1007/s10922-020-09564-7.

[5] Dutch National Internet Providers Management Organization, "The state of affairs regarding the recent DDoS attacks," September 2020.

[6] Henry Farrell and Abraham L. Newman, "Weaponized Interdependence," *International Security*, vol. 44, no. 1, 2019.

[7] Big Data Innovation Hubs with Trusted CI, "Cybersecurity as Big Data Science Workshop," April 2021. https://nebigdatahub.org/cybersecurity-as-big-data-science-workshop/.

[8] The Networking and Information Technology R&D Program Large Scale Networking Interagency Working Group, "Huge Data: A Computing, Networking, and Distributed Systems Perspective Workshop Report," January 2021.

[9] Center for Applied Internet Data Analysis, "Active Internet Measurement Systems." https://www.caida.org/workshops/?workshopserieslisting=AIMS.

[10] k. claffy, M. Fomenkov, E. Katz-Bassett, R. Beverly, B. Cox, and M. Luckie, "The Workshop on Active Internet Measurements (AIMS) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 39, Oct 2009.

[11] kc claffy, E. Aben, J. Augé, R. Beverly, F. Bustamante, B. Donnet, T. Friedman, M. Fomenkov, P. Haga, M. Luckie, and Y. Shavitt, "The 2nd Workshop on Active Internet Measurements (AIMS-2) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 40, Oct. 2010.

[12] kc claffy, "The 3rd Workshop on Active Internet Measurements (AIMS-3) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 41, July 2011.

[13] kc claffy, "The 4th Workshop on Active Internet Measurements (AIMS-4) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 42, Jul 2012.

[14] kc claffy, "The 5th Workshop on Active Internet Measurements (AIMS-5) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 43, Jul 2013.

[15] kc claffy, "The 6th Workshop on Active Internet Measurements (AIMS-6) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 44, Oct 2014.

[16] kc claffy, "The 7th Workshop on Active Internet Measurements (AIMS-7) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 46, Jan 2015. http://www.caida.org/publications/papers/2016/aims2015_report/.

[17] kc claffy, "The 8th Workshop on Active Internet Measurements (AIMS-8) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Oct 2016. http://www.caida.org/publications/papers/2016/aims2016_report/.

[18] kc claffy, "The 9th Workshop on Active Internet Measurements (AIMS-9) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Oct 2017. http://www.caida.org/publications/papers/2017/aims2017_report/.

[19] kc claffy, "The 10th Workshop on Active Internet Measurements (AIMS-10) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Oct 2018. http://www.caida.org/publications/papers/2018/aims2018_report/.

[20] kc claffy, "The 11th Workshop on Active Internet Measurements (AIMS-11) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, forthcoming. http://www.caida.org/publications/papers/2019/aims2019_report/.

[21] Center for Applied Internet Data Analysis, "Workshop on Internet Economics." https://www.caida.org/workshops/?workshopserieslisting=WIE.

[22] k. claffy, "Workshop on Internet Economics (WIE2011) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 40, Apr 2010.

[23] k. claffy, "Workshop on Internet Economics (WIE2011) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 42, pp. 110–114, Apr 2012.

[24] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2012) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 43, pp. 95–100, Jul 2013.

[25] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2013) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 44, pp. 116–119, Jul 2014.

[26] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2014) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 45, pp. 43–48, Jul 2015.

[27] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2015) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jul 2016.

[28] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2016) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jul 2017.

[29] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2017) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jul 2018.

[30] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2018) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jan 2019.

[31] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2019) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jan 2020.

[32] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2020) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jan 2020.

[33] Serge Fdida and Ivan Seskar and Peter Steenkiste and Brecht Vermeulen, "Report from the EU/US Future Networks Workshop," November 2017. https://www.cs.cmu.edu/~prs/eu-us-report-nov-2017.pdf.

[34] Center for Applied Internet Data Analysis, "Workshop on Overcoming Barriers to Internet Research." https://www.caida.org/workshops/?workshopserieslisting=WOMBIR.

[35] A. Marder, M. Luckie, A. Dhamdhere, B. Huffaker, J. Smith, and k. claffy, "Pushing the boundaries with bdrmapit: Mapping router ownership at internet scale," in *ACM Internet Measurement Conference (IMC)*, pp. 56–69, 11 2018.

[36] A. Shah, R. Fontugne, and C. Papadopoulos, "Router-level topologies of autonomous systems," in *Complex Networks*, 2018.

[37] J. Hu, Z. Zhou, X. Yang, J. Malone, and J. W. Williams, "Cablemon: Improving the reliability of cable broadband networks via proactive network maintenance," in *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, (Santa Clara, CA), pp. 619–632, USENIX Association, Feb. 2020.

[38] Z. S. Bischof, R. Fontugne, and F. E. Bustamante, "Untangling the world-wide mesh of undersea cables," in *Proceedings of the 17th ACM Workshop on Hot Topics in Networks*, HotNets '18, (New York, NY, USA), p. 7884, Association for Computing Machinery, 2018.

[39] R. Fanou, B. Huffaker, R. Mok, and k. claffy, "Unintended consequences: Effects of submarine cable deployment on Internet routing," in *Passive and Active Measurement Conference (PAM)*, Mar 2020.

[40] K. G. Leyba, B. Edwards, C. Freeman, J. R. Crandall, and S. Forrest, "Borders and Gateways: Measuring and Analyzing National AS Chokepoints," in *COMPASS*, 2019.

[41] A. Shah, R. Fontugne, and C. Papadopoulos, "Towards characterizing international routing detours," in *Asian Internet Engineering Conference (AINTEC)*, 2016.

[42] t. Freyburg and L. Garbe, "Blocking the bottleneck: Internet shutdowns and ownership at election times in sub-saharan africa," in *International Journal of Communication*, 2018.

[43] A. Lutu, B. Jun, F. E. Bustamante, D. Perino, M. Bagnulo, and C. G. Bontje, "A first look at the ip exchange ecosystem," *SIGCOMM Comput. Commun. Rev.*, vol. 50, p. 2534, Oct. 2020.

[44] G. Ramezan, C. Leung, and Z. J. Wang, "A survey of secure routing protocols in multi-hop cellular networks," in *IEEE Communications Survey & Tutorials*, 2018.

[45] R. Boutaba, M. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O. Caicedo Rendon, "A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities," *Journal of Internet Services and Applications*, vol. 9, 05 2018.

[46] Y. Lavinia, R. Durairajan, R. Rejaie, and W. Willinger, "Challenges in using ml for networking research: How to label if you must," in *Proceedings of the Workshop on Network Meets AI and ML*, NetAI '20, (New York, NY, USA), p. 2127, Association for Computing Machinery, 2020.

[47] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data management challenges in production machine learning," in *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, (New York, NY, USA), p. 17231726, Association for Computing Machinery, 2017.

[48] H. Hindy, D. Brosset, E. Bayne, A. K. Seeam, C. Tachtatzis, R. Atkinson, and X. Bellekens, "A taxonomy of network threats and the effect of current datasets on intrusion detection systems," *IEEE Access*, vol. 8, pp. 104650–104675, 2020.

[49] J. Postel, "Internet Protocol." RFC 791, Sept. 1981.

[50] Y. Rekhter (Ed.), T. Li (Ed.), and S. Hares (Ed.), "A Border Gateway Protocol 4 (BGP-4)." RFC 4271 (Draft Standard), Jan. 2006. Updated by RFCs 6286, 6608, 6793, 7606, 7607, 7705, 8212, 8654.

[51] D. Eastlake and C. Kaufman, "Domain name system security extensions." RFC 2065, Jan. 1997.

[52] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose, "DNS security introduction and requirements." RFC 4033, Mar. 2005.

[53] P. Ferguson and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks Which Employ IP Source Address Spoofing." RFC 2827, May 2000.

[54] H. Ballani, P. Francis, and X. Zhang, "A study of prefix hijacking and interception in the internet," in *Proceedings of the 2007 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '07, (New York, NY, USA), p. 265276, Association for Computing Machinery, 2007.

[55] Arstechnica, "Another fraudulent certificate raises the same old questions about certificate authorities," Aug. 2011.

[56] K. Man, Z. Qian, Z. Wang, X. Zheng, Y. Huang, and H. Duan, "Dns cache poisoning attack reloaded: Revolutions with side channels," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, (New York, NY, USA), p. 13371350, Association for Computing Machinery, 2020.

[57] M. Jonker, A. King, J. Krupp, C. Rossow, A. Sperotto, and A. Dainotti, "Millions of targets under attack: a macroscopic characterization of the DoS ecosystem," in *IMC*, Nov. 2017.

[58] S. Kottler, "February 28th DDoS Incident Report." `https://github.blog/2018-03-01-ddos-incident-report/`.

[59] M. Luckie, R. Beverly, R. Koga, K. Keys, J. Kroll, and k. claffy, "Network Hygiene, Incentives, and Regulation: Deployment of Source Address Validation in the Internet," in *ACM Computer and Communications Security (CCS)*, Nov 2019.

[60] B. Krebs, "The democratization of censorship," Sept. 2016. `https://krebsonsecurity.com/2016/09/the-democratization-of-censorship/`.

[61] Arbor Networks, "Worldwide infrastructure security report," 2018. `https://resources.arbornetworks.com/`.

[62] L. Prehn, F. Lichtblau, and A. Feldmann, "When wells run dry: The 2020 ipv4 address market," in *Proceedings of the 16th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '20, (New York, NY, USA), p. 4654, Association for Computing Machinery, 2020.

[63] Y. Sun, M. Apostolaki, H. Birge-Lee, L. Vanbever, J. Rexford, M. Chiang, and P. Mittal, "Securing internet applications from routing attacks," 2020.

[64] C. Orsini, A. King, D. Giordano, V. Giotsas, and A. Dainotti, "BGPStream: a software framework for live and historical BGP data analysis," in *Internet Measurement Conference (IMC)*, Nov 2016.

[65] Y.-J. Chi, R. Oliveira, and L. Zhang, "Cyclops: The as-level connectivity observatory," *SIGCOMM Comput. Commun. Rev.*, vol. 38, p. 516, Sept. 2008.

[66] J. Schlamp, R. Holz, Q. Jacquemart, G. Carle, and E. W. Biersack, "Heap: reliable assessment of bgp hijacking attacks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 6, pp. 1849–1861, 2016.

[67] M. Lad, D. Massey, D. Pei, Y. Wu, B. Zhang, and L. Zhang, "Phas: A prefix hijack alert system.," in *USENIX Security symposium*, vol. 1, p. 3, 2006.

[68] G. Chaviaras, P. Gigis, P. Sermpezis, and X. Dimitropoulos, "Artemis: Real-time detection and automatic mitigation for bgp prefix hijacking," in *Proceedings of the 2016 ACM SIGCOMM Conference*, pp. 625–626, 2016.

[69] B. Al-Musawi, P. Branch, and G. Armitage, "Bgp anomaly detection techniques: A survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. 377–396, 2017.

[70] T. Hlavacek, I. Cunha, Y. Gilad, A. Herzberg, E. Katz-Bassett, M. Schapira, and H. Shulman, "Disco: Sidestepping rpki's deployment barriers," 01 2020.

[71] Z. Zhang, Y. Zhang, Y. C. Hu, Z. M. Mao, and R. Bush, "ispy: Detecting ip prefix hijacking on my own," in *Proceedings of the ACM SIGCOMM 2008 conference on Data Communication*, pp. 327–338, 2008.

[72] L. Peterson, T. Anderson, D. Culler, and T. Roscoe, "A blueprint for introducing disruptive technology into the Internet," in *Hotnets*, 2002.

[73] E. Katz-Bassett, H. V. Madhyastha, J. P. John, A. Krishnamurthy, D. Wetherall, and T. E. Anderson, "Studying black holes in the internet with hubble.," in *NSDI*, vol. 8, pp. 247–262, 2008.

[74] X. Shi, Y. Xiang, Z. Wang, X. Yin, and J. Wu, "Detecting prefix hijackings in the internet with argus," in *Proceedings of the 2012 Internet Measurement Conference*, pp. 15–28, 2012.

[75] X. Hu and Z. M. Mao, "Accurate real-time identification of ip prefix hijacking," in *2007 IEEE Symposium on Security and Privacy (SP'07)*, pp. 3–17, IEEE, 2007.

[76] C. Testart, P. Richter, A. King, A. Dainotti, and D. Clark, "Profiling BGP Serial Hijackers: Capturing Persistent Misbehavior in the Global Routing Table," in *ACM Internet Measurement Conference (IMC)*, Oct 2019.

[77] M. Luckie, B. Huffaker, A. Dhamdhere, V. Giotsas, and k. claffy, "AS Relationships, Customer Cones, and Validation," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, Oct 2013.

[78] M. Konte, R. Perdisci, and N. Feamster, "Aswatch: An as reputation system to expose bulletproof hosting ases," *SIGCOMM Comput. Commun. Rev.*, vol. 45, p. 625638, Aug. 2015.

[79] C. Testart and D. Clark, "A Data-Driven Approach to Understanding the State of Internet Routing Security," in *Telecommunications Policy Research Conference (TPRC)*, Feb 2021.

[80] Internet Society, "Mutually Agreed Norms for Routing Security." `https://www.manrs.org/`.

[81] Lehr, W. and Clark, D., "Changing Markets for Domain Names: Technical, Economic, and Policy Challenges," in *Telecommunications Policy Research Conference (TPRC)*, Feb 2021. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3746594`.

[82] A. Kashaf, V. Sekar, and Y. Agarwal, "Analyzing third party service dependencies in modern web services: Have we learned from the mirai-dyn incident?," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), p. 634647, Association for Computing Machinery, 2020.

[83] E. Alowaisheq, S. Tang, Z. Wang, F. Alharbi, X. Liao, and X. Wang, "Zombie awakening: Stealthy hijacking of active domains through dns hosting referral," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, (New York, NY, USA), p. 13071322, Association for Computing Machinery, 2020.

[84] M. Allman, "Comments on dns robustness," in *Proceedings of the Internet Measurement Conference 2018*, IMC '18, (New York, NY, USA), p. 8490, Association for Computing Machinery, 2018.

[85] G. Akiwate, M. Jonker, R. Sommese, I. Foster, G. M. Voelker, S. Savage, and K. Claffy, "Unresolved Issues: Prevalence, Persistence, and Perils of Lame Delegations," in *Proceedings of the ACM Internet Measurement Conference (IMC)*, (Virtual Event), October 2020.

[86] A. Oest, P. Zhang, B. Wardman, E. Nunes, J. Burgis, A. Zand, K. Thomas, A. Doupé, and G.-J. Ahn, "Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 361–377, USENIX Association, Aug. 2020.

[87] B. Zhao, M. Ikram, H. Asghar, M. A. Kaafar, A. Chaabane, and K. Thilakarathna, "A decade of mal-activity reporting: A retrospective analysis of internet malicious activity blacklists," pp. 193–205, 07 2019.

[88] S. Tajalizadehkhoob, R. Böhme, C. Gañán, M. Korczyński, and M. Van Eeten, "Rotten Apples or Bad Harvest? What We Are Measuring When We Are Measuring Abuse," *arXiv e-prints*, p. arXiv:1702.01624, Feb. 2017.

[89] V. G. Li, M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, and S. Savage, "Reading the tea leaves: A comparative analysis of threat intelligence," in *28th USENIX Security Symposium (USENIX Security 19)*, (Santa Clara, CA), pp. 851–867, USENIX Association, Aug. 2019.

[90] D. Liu, Z. Li, K. Du, H. Wang, B. Liu, and H. Duan, "Don't let one rotten apple spoil the whole barrel: Towards automated detection of shadowed domains," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pp. 537–552, ACM, 2017.

[91] Thomas Vissers and Jan Spooren and Pieter Agten and Dirk Jumpertz and Peter Janssen and Marc Van Wesemael and Frank Piessens and Wouter Joosen and Lieven Desmet, "Exploring the Ecosystem of Malicious Domain Registrations in the .eu TLD," 2017. https://link.springer.com/chapter/10.1007/978-3-319-66332-6_21.

[92] N. Serrano, H. Hadan, and L. Camp, "A complete study of p.k.i. (pkis known incidents)," *SSRN Electronic Journal*, 01 2019.

[93] Q. Scheitle, O. Gasser, T. Nolte, J. Amann, L. Brent, G. Carle, R. Holz, T. C. Schmidt, and M. Wählisch, "The rise of certificate transparency and its implications on the internet ecosystem," in *Proceedings of the Internet Measurement Conference 2018*, IMC '18, (New York, NY, USA), p. 343349, Association for Computing Machinery, 2018.

[94] Pouyan Fotouhi Tehrani and Eric Osterweil and Jochen H. Schiller and Thomas C. Schmidt and Matthias Whlisch, "Security of Alerting Authorities in the WWW: Measuring Namespaces, DNSSEC, and Web PKI," in *Proceedings of the Web Conference 2021 (WWW '21)*, 2021. https://www2021.thewebconf.org/papers/security-of-alerting-authorities-in-the-www-measuring-namespaces-dnssec-and-w

[95] Zane Ma and Joshua Mason and Manos Antonakakis and Zakir Durumeric and Michael Bailey, "What's in a name? exploring CA certificate control," in *30th USENIX Security Symposium (USENIX Security 21)*, USENIX Association, Aug. 2021.

[96] D. Clark and k. claffy, "Toward a Theory of Harms in the Internet Ecosystem," in *Telecommunications Policy Research Conference (TPRC)*, Sep 2019.

[97] "Online Harms White Paper: Full Government Response to the consultation," December 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/944310/Online_Harms_White_Paper_Full_Government_Response_to_the_consultation_CP_354_CCS001_CCS1220695430-001__V2.pdf.

[98] Apache Software Foundation, "Kafka," 2021. https://kafka.apache.org.

[99] Lepinski, Matt and Kent, Stephen and Kong, Derrick, "A profile for route origin authorizations (ROAs)," *IETF, RFC*, vol. 6482, 2012.

[100] T. Hlavacek, I. Cunha, Y. Gilad, A. Herzberg, E. Katz-Bassett, M. Schapira, and H. Shulman, "DISCO: Sidestepping RPKI's Deployment Barriers," in *Proceedings 2020 Network and Distributed System Security Symposium*, (San Diego, CA), Internet Society, 2020.

[101] K. Benson, A. Dainotti, k. claffy, A. Snoeren, and M. Kallitsis, "Leveraging internet background radiation for opportunistic network analysis," in *ACM Internet Measurement Conference (IMC)*, 10 2015.

[102] National Science Foundation and National Radio Astronomy Observatory, "Very Long Baseline Array Telescope." https://public.nrao.edu/telescopes/vlba/.

[103] Matthew Luckie, Ken Keys, Ryan Koga, Rob Beverly, kc claffy, "Spoofer source address validation measurement system," 2016. http://spoofer.caida.org.

[104] G. Wan, L. Izhikevich, D. Adrian, K. Yoshioka, R. Holz, C. Rossow, and Z. Durumeric, "On the origin of scanning: The impact of location on internet-wide scans," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), p. 662679, Association for Computing Machinery, 2020.

[105] B. Tierney, J. Boote, E. Boyd, A. Brown, M. Grigoriev, J. Metzger, M. Swany, M. Zekauskas, Y.-T. Li, and J. Zurawski, "Instantiating a Global Network Measurement Framework," Tech. Rep. LBNL-1452E, LBNL, Jan. 2009. http://acs.lbl.gov/~tierney/papers/perfsonar-LBNL-report.pdf.

[106] P. Sermpezis, V. Kotronis, P. Gigis, X. Dimitropoulos, D. Cicalese, A. King, and A. Dainotti, "ARTEMIS: Neutralizing BGP Hijacking within a Minute," *IEEE/ACM Transactions on Networking*, vol. 26, pp. 2471–2486, Dec 2018.

[107] K. Levchenko, A. Dhamdhere, B. Huffaker, k. claffy, M. Allman, and V. Paxson, "PacketLab: A Universal Measurement Endpoint Interface," in *ACM Internet Measurement Conference (IMC)*, Nov 2017. https://packetlab.github.io/.

[108] GEANT Consoritum and European Commission, "Next Generation GEANT backbone," 2020. https://www.geant.org/Projects/GEANT_Project_GN4-3/Pages/GN4-3N.aspx.

[109] Joe Breen and Dan Doyle and James Deaton, "Gathering Basic Metrics and Telemetry - Community Efforts," February 2021. Internet2 Performance Working Group Community Metric and Telemetry Project Team, slide presentation available on request.

[110] Emile Aben, David Teach, Bradley Huffaker and kc claffy, "Route Collector MetaData: Document for Community Feedback," 2021. https://github.com/CAIDA/route-collectors/wiki/RouteCollectionMetaData.

[111] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daum III, and K. Crawford, "Datasheets for datasets," tech. rep., Microsoft, March 2018.

[112] Joel Sommers, "Automating Active Measurement Metadata Collection and Analysis," 2021. https://cs.colgate.edu/~jsommers/someta.html.

[113] Wilkinson, Mark D. et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Nature Scientific Data*, 2016. https://doi.org/10.1038/sdata.2016.18.

[114] Julia Lane and Paco Nathan, "Rich Context Workshop," 2020. https://coleridgeinitiative.org/richcontext/richcontextworkshop/.

[115] Bradley Huffaker and kc claffy and Elena Yulaeva and Josh Polterock and Jon Weber, "CAIDA's Internet Resource Catalog," 2020.

[116] Yuanlong Tan and Malathi Veeraraghavan and Hwajung Lee and Steve Emmerson and Jack Davidson, "A Trial Deployment of a Reliable Network-Multicast Application across Internet2," 2020.

[117] Christine R. Kirkpatrick and Kevin Coakley and Melissa Cragin and James Glasgow and John Goodhue, "Research Drivers and Capabilities - OpenStorageNetwork Concept Paper," December 2020. https://www.openstoragenetwork.org/wp-content/uploads/2020/12/Research-Drivers-and-Capabilities-Concept-Paper-2.pdf.

[118] Ken Keys, "Periscope Looking Glass API." https://www.caida.org/tools/utilities/looking-glass-api/.

[119] Roland van Rijswijk-Deij, "OpenINTEL update." https://ripe78.ripe.net/presentations/4-20190520-RIPE-78-Reykjavik-OpenINTEL.pdf.

[120] Ian Foster and CAIDA, "DNS TLD Zone DB," 2020. https://dzdb.caida.org.

[121] REN-ISAC, "What is Passive DNS." https://www.ren-isac.net/member-resources/pDNS.html.

[122] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar, "Using gans for sharing networked time series data: Challenges, initial promise, and open questions," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), p. 464483, Association for Computing Machinery, 2020.

[123] C. Staff, "Differential privacy: The pursuit of protections by default," *Commun. ACM*, vol. 64, p. 3643, Jan. 2021.

[124] V. Dani, V. King, M. Movahedi, J. Saia, and M. Zamani, "Secure Multi-Party Computation in Large Networks," *arXiv e-prints*, p. arXiv:1203.0289, Mar. 2012.

[125] Y. Lindell, "Secure multiparty computation," *Commun. ACM*, vol. 64, p. 8696, Dec. 2020.
[126] P. Francis, S. Eide, and R. Munz, "Diffix: High-utility database anonymization," in *Annual Privacy Forum*, pp. 141–158, 06 2017.
[127] H. Corrigan-Gibbs and D. Boneh, "Prio: Private, robust, and scalable computation of aggregate statistics," in *Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation*, NSDI'17, (USA), p. 259282, USENIX Association, 2017.
[128] D. of Homeland Security, "PREDICT project: Protected Repository for Defense of Infrastructure against Cyber Threats." `http://www.predict.org/`.
[129] Department of Homeland Security, "Information Market for Policy and Analysis of Cyberrisk and Trust," 2020. `https://www.impactcybertrust.org/`.
[130] Kenneally, Erin and Dittrich, David, "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research," 2012. `http://ssrn.com/abstract=2445102`.
[131] Dittrich, David and Kenneally, Erin and Bailey, Michael, "Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Menlo Report," 2013. `http://ssrn.com/abstract=2342036`.
[132] Aaron Bursteion, "Amending the ECPA to enable a culture of cybersecurity research," 2008. `http://jolt.law.harvard.edu/articles/pdf/v22/22HarvJLTech167.pdf`.
[133] S. J. Saidi, A. M. Mandalari, R. Kolcun, H. Haddadi, D. J. Dubois, D. Choffnes, G. Smaragdakis, and A. Feldmann, "A haystack full of needles: Scalable detection of iot devices in the wild," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), p. 87100, Association for Computing Machinery, 2020.