Workshop on Internet Economics (WIE 2019) report

kc claffy
UCSD/CAIDA
kc@caida.org

David Clark
MIT/CSAIL
ddc@csail.mit.edu

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

On 9-11 December 2019, CAIDA hosted the 10th interdisciplinary Workshop on Internet Economics (WIE) at UC San Diego's Supercomputer Center. This workshop series provides a forum for researchers, Internet facilities and service providers, technologists, economists, theorists, policymakers, and other stakeholders to exchange views on current and emerging economic and policy debates. This year's meeting had a narrower focus than in years past, motivated by a new NSF-funded project being launched at CAIDA: KISMET (Knowledge of Internet Structure: Measurement, Epistemology, and Technology). The objective of the KISMET project is to improve the security and resilience of key Internet systems by collecting and curating infrastructure data in a form that facilitates query, integration and analysis. This project is a part of NSF's new Convergence Accelerator program, which seeks to support fundamental scientific exploration by creating partnerships across public and private sectors to solve problems of national importance.

1 MOTIVATION

In the mid-1990s, the U.S. government launched strategic industrial policies to promote competition, and thus innovation, in the emerging Internet transport and domain name industries. In the 25 years since, the Internet's reach has expanded to over 3B people around the world, and continues to grow. As the Internet has become critical infrastructure, society has grown increasingly exposed to its security weaknesses. Despite extensive efforts across industry, government, NGOs, and academia to mitigate many of these weaknesses, security issues continue to capture daily headlines. Call for regulatory oversight in the face of professed market failures beg several empirical questions. What are the biggest security threats to the Internet infrastructure? How can we understand, if not quantify, the effectiveness of risk-mitigating efforts, or even to what extent defenses have been deployed?

We focus on the three fundamental layers of the Internet architecture that require some level of global governance to guarantee consistent and reliable interpretation: addressing, routing, and naming. The Internet routing and naming ecosystems are both characterized by dynamics that have no rigorous theoretical foundation and radically distributed ownership. Furthermore, operational practices surrounding these layers have enabled malicious actors to successfully execute, and scale, harmful misbehavior, including DoS and phishing attacks as well as distribution and execution of sophisticated malware [10]. Their complexity means that distinguishing malicious behavior from sophisticated network engineering is a grand challenge of 21st century Internet science.

The epistemological challenges span many disciplines. Engineering, science, economics, and public policy communities have long struggled to understand aspects of the global Internet ecosystem. Although many data sources exist in various forms, the volume and complexity of data is overwhelming. Turning data into knowledge is a challenge that requires multidisciplinary investment, including network operations expertise, data science methods, systems integration and administration effort, and legal knowledge to inform data sharing risk assessments and disclosure controls.

We used this workshop as a forum to explore the feasibility and long term sustainability of an Open Knowledge Network (OKN) of public data on Internet structure, as manifested in the naming, addressing, and routing systems, to confront the growing empirical gap in science, security, and public communications policy. The hypothesis is that higher quality and more accessible data will enable better decision-making, direction-setting and improvement in Internet security and resilience. We organized the workshop around the application of five related questions to each subsystem.

- Can we taxonomize the threats to each subsystem?
- What relevant datasets are available for scientific research?
- What additional data would facilitate threat analysis?
- Who collects data, and how do they support the collection?
- What institutional or organizational gaps exist?

We organize this report into descriptions of the threats at different layers, proposed research directions to mitigate them, and existing and potential data sets as well as potential business models to sustain their production and sharing in the context of an OKN.

2 THREATS

Addressing. Routers forwarding Internet packets do not generally examine source addresses when making forwarding decisions. The specification of the Internet Protocol (IP) does not require verification of the source addresses in packets. This design decision renders it possible to use spoofed source addresses, and malicious actors exploit this spoofing ability to execute a wide variety of attacks [25]. The IETF has recommended source address validation (SAV) as a best practice for over a decade [3, 13]. But networks have little incentive to comply since proper filtering requires expertise and maintenance, and only helps other networks. Networks that allow spoofing reduce their own operational costs, while imposing costs on others, in the form of attacks and attack risk.

The measurement challenge makes SAV particularly intractable. Unlike many security best practices that can be measured from anywhere on the network, measurement of SAV on a network requires attempting to transmit an invalid-source addressed packet

from that network to the public Internet. Any regulatory, procurement, insurance, or peering requirement for SAV compliance would require this measurement capability. Since there will always be countries who do not establish or enforce such requirements, others have called for an application-layer solutions to limiting the amplification power of the spoofing vector [32], but such solutions cannot prevent other harms due to spoofing [23, 27].

An additional authentication issue is the lack of a trustworthy registry of data on which organizations has operational authority over which IP addresses. Each of the five Regional Internet Registries (RIRs) executes its own multistakeholder policy to coordinate address allocation within its region. There are historic address allocation and assignment databases at each RIR, but IPv4 address blocks can now be bought, sold, and leased, with varying degrees of transparency, including across RIR regions. The RIRs do not operate registries of operational control or intended routing policies.

Interdomain routing system. BGP is the global routing protocol of the Internet, by which autonomous networks (autonomous systems or ASes) propagate network topology information. A variety of operational disruptions derive from mis-announcement of BGP prefixes [34]. For over two decades the Internet engineering community has tried to develop and deploy interdomain routing security mechanisms, with little success. One such technology recently gaining some traction is the use of *Route Origin Authorizations (ROAs).* The RIRs and the IETF developed a protocol for voluntary use of ROAs to establish definitive authority to originate a specified prefix into the interdomain routing system. If pervasively used, this mechanism can minimize the threat of certain forms of hijacking (including accidental) [34]. Uptake of this technology is low, although growing [8].

According to data from NIST, as of January 2020, about 19% of the routed address space is covered by ROAs globally [31]. The coverage differs by region: in Europe, it is close to 49%; in the Americas (ARIN) it is 9%. One reason for low registration in the ARIN region is the terms of the contract with ARIN required to register ROAs. Christopher Yoo (UPenn) presented his efforts to analyze these contracts in pursuit of making them less burdensome. When ROAs were first deployed, U.S. ISPs were concerned about the risk of inconsistency between registrations and actual routing assertions, which would trigger dropping of routes. However, the NIST data shows that currently only .3% of the routed address space is represented by routing assertions that are invalid based on a ROA.

In the face of the political/economic challenges in incentivizing deployment of ROAs thus far, some have argued for the use of open BGP monitoring platforms and routing policy registries to enable identification of anomalies from a *baseline* BGP state. Others point out the inherent difficulty of distinguishing anomalous BGP states intended to harm or exploit, e.g., hijacks, from complex routing policies to support traffic engineering, e.g, Figure 7 of [26]. The vibrant but opaque political economy in which the routing system operates presents challenges in how to think about a knowledge network to support routing infrastructure security.

Naming System. There is increasing concern that the multistakeholder governance model of the DNS ecosystem has created a haven for criminal and fraudulent activities. In January 2019, the U.S. DHS issued emergency directive 19-01 [36], requiring government agencies to implement best practices to protect DNS infrastructure from attacks. This directive adds to the hundreds of pages of recommended cybersecurity best practices by various U.S. agencies.

There are two primary classes of threats to the naming system. First, the DNS can generate false mappings of names to IP addresses, due to lack of authentication in the DNS transaction. Cryptographic DNS zone signing technology has been around for over a decade but is still not well used. Second, the name registration ecosystem supports extremely opaque use of names, in the interests of privacy, but also unfortunately in the interest of malicious actors.

Competitive pressures inhibit investment in security measures, and the self-regulatory model of governance of Internet identifiers is struggling to achieve its own standards for accountability [5]. Europe's recent launch of the General Data Protection Regulation has prompted ICANN's greatest challenge yet, a conflict unresolved for decades over what metadata about Internet identifier ownership should be available to whom. This meta data is a pillar in operational security efforts to combat cybercrime and other malfeasance rooted in inappropriate use of Internet names and numbers and multiple multistakeholder processes are attempting to expeditiously reach compromise [20, 22]. One stakeholder absent from these conversations is the scientific research community, both from the academic and government sectors.

3 DATA SOURCES TO SUPPORT AN OPEN KNOWLEDGE NETWORK

At the workshop we discussed several raw and derivative data sets that are promising components of an open knowledge network on Internet structure. Despite the large swath of data being collected and shared, data collection is driven more by operational than scientific need, and in both cases is constrained by the practicality and cost of collecting it, which often limits its utility.

3.1 Addressing and routing data

Address ownership. IANA maintains and publishes data of IP address delegations to regional registries, which in turn maintain and publish information about subsequent allocation and assignment of addresses to either local Internet registries or enterprises.

Routing topology. There are independent routing policy registries where networks may voluntarily register their own policy information[1], but there is not a centralized form of such data, and these assertions may not match actual route configurations. Thus, interdomain routing analysis largely relies on instrumentation that participates in the Border Gateway Protocol (BGP) to receive (and thus observe) reachability information from routers. Two organizations – U. Oregon and RIPE RIS – collect and store such Internet interdomain routing data from several locations around the globe. The Network Startup Resource Center (NSRC) at the University of Oregon operates RouteViews, originally conceived as a tool for Internet operators to obtain real-time BGP information about how others viewed their prefixes. NSRC is planning the next generation of BGP data collection and distribution infrastructure, which will

[1] http://www.irr.net/docs/list.html

include per-message timestamps, meta-data, real-time streaming telemetry using the OpenBMP protocol, automated consolidation and sequencing, and RPKI validation and retrieval.

Another BGP measurement project, bgpmon.net, started by an individual (Andre Toonk), allows users to monitor changes in observed announcements of specific prefixes for free (an example of turning data into useful knowledge). In 2012, to sustain the service, he commercialized it. In 2015, OpenDNS acquired BGPmon, and later that year Cisco acquired OpenDNS [9]. In 2018 Cisco announced plans to convert this service into a proprietary cloud service, but include a limited free alert service at bgpstream.com of potential route leaks based on its set of peers.

Some workshop participants found there to be an overwhelming amount of raw BGP data, but a paucity of derivative and curated data sets (and improved geolocation) that facilitate security-related research on the data.

The Internet Society's Mutually Agreed Norms for Routing Security (MANRS) project defines actions that ISPs should undertake to reduce threats such as route hijacks and leaks. The Internet Society does not itself gather data to evaluate compliance by its members with these expectations, nor study the effectiveness of these actions in reducing threats. They currently use alerts generated by Cisco's bgpstream.com to populate its MANRS observatory, but are eager to have a higher fidelity source of data.

Finally, RIPE archives the historical record of ROAs for all RIRs since 2011. As an example of turning raw data into open and security-relevant knowledge, RIPE maintains an archive of all ROAs from the five RIRs on a daily basis. RIPE also provides a curated data set where they have validated the ROAs for each day, and compared them to BGP routing announcements observed at the corresponding time. This curated form facilitates analysis of *ROA-invalid* announcements, and which ASes generated them.

Network/organizational structure. There is no official database mapping AS numbers to organizations owning them; CAIDA maintains a heuristic-based mapping of ASes to organizations [18] as well as a mapping of which IP addresses (v4 and v6) each network originates into the global routing system [7].

Internet Topology Data Kits (ITDKs). Using raw traceroute and BGP data, CAIDA publishes derivative data sets, including heavily curated two-week snapshots of raw traceroute data into Internet Topology Data Kits (ITDK) [6]. Each ITDK contains inferred, DNS-annotated, router-level and AS-level topologies of a large cross-section of the global Internet.

Inferred Relationships Between Networks. CAIDA operates a service (AS Rank) for exploration of routing and business relationships between ASes and organizations that own them.

Geolocation data. There is no universal public database of geolocation of Internet identifiers; several commercial companies sell access to proprietary databases for IP address geolocation, and some have also supported research use of the data. Researchers have compared accuracy across databases, and found that some databases are reasonably accurate for edge (e.g., host) infrastructure, but geolocation of core (router) infrastructure is a harder challenge (and/or lower priority) [14].

3.2 Domain name system (DNS) data

The DNS is a globally distributed database that maps domain names (e.g., www.example.com) to IP addresses. Data collected about the DNS may facilitate operational and academic research, as well as law enforcement, e.g., to combat phishing, spam, brand infringements, and other malicious uses of domains. The DNS involves more types of players in the ecosystem than BGP, yielding many options for data collection. Users ("registrants") can purchase domain names, which requires information (sometimes false or withheld from public view) about the owners of those names.

Absent from this list is financial data. Domain names sell at widely varying prices by different registries and registrars. Names that sell at a low price are, not surprisingly, appealing to miscreants who essentially need throw-away or burner names. Some registrars have sales, i.e., periods when they sell names for a few cents, triggering bulk purchases that often suggest malicious intent. Much data about pricing and business terms of domain name acquisition is private, although there are known strong correlations between low pricing and malicious domain registrations [16]. This issue was articulated by multistakeholder review teams of ICANN's performance in the areas of security [33] and consumer trust [11].

ICANN Centralized Zone Data Service. Each Top Level Domain (TLD) registry operator maintains a zone file that contains information on domains, including associated name server hosts, and IP addresses for those name servers. TLD zone data is inherently public via DNS queries but acquiring an entire zone file for research has historically required applying for access from each TLD registry operator, under appropriate use terms, e.g., no spamming of domains. In 2012, ICANN established a centralized data access platform to simplify access to all zone files for new gTLDs. Registry operators are required to upload a copy of their current zones to ICANN every 24 hours. (Legacy gTLDs and country-code TLDs, are not contractually required to participate in this program.) Individuals and organizations may collect and archive these and other zone files over time, and even create other services based on this data. As an example, the DNSCoffee [19] project has downloaded CZDS zone files as well as other zone files for several years. Ian Foster maintained this project as a UCSD student to support his research [16, 17]. Of those organizations represented at the workshop, Farsight (commercial), DNS-OARC, and Interisle also gather some subset of zone files.

Active measurements of DNS namespace. NLNet labs (LoC) – jointly with three Dutch research institutions (SURFnet, SIDN Labs, and U. Twente) – operates the OpenINTEL project (started in 2015), a system for comprehensive active measurements of the global DNS [37]. OpenINTEL uses ICANN's CZDS files, and agreements with many other registries, to drive DNS queries for all covered domains once every 24 hours, covering over 220 million domains per day for: .com, .net, .org, .info, .mobi, .aero, .asia, .name, .biz, .gov, almost 1200 new gTLDs and many ccTLDs. They have used this data to study and improve DNSSEC operational practices, DNS resilience, and identify misconfigurations.

"Passive DNS" (above recursive). Several companies collect queries sent by recursive resolvers to authoritative resolvers. As an example, Farsight has been collecting and selling access to raw

and derivative forms of this data for years—they state that they hold over 100B DNS records, captured from actual queries.

Passive DNS (below resolvers). Several participants noted the potential benefit of gathering queries from end clients *to* recursive resolvers, to identify, for example, what fraction of queries by users were to known legitimate domains. Such queries will have user IP addresses that are generally considered Personally Identifying Information (PII) so would require some appropriate disclosure control approach to sharing. The complexity of collection and sharing and lack of acute need has kept this type of data set from coming into use by researchers.

Root server packet traces. OARC captures the *Day in the Life* (DITL) collection, which is an annual, one day collection of queries to a number of busy authoritative servers, mostly root servers. DITL surveys have been conducted since 2006. OARC is a small organization (less than 4 FTEs) and does not have the resources to track what scientific studies are published using the data.

Registration information on second-level zones. Zone files show when a new name has been registered, and a researcher can query the registry to learn some information about the registration. However, the available information has dropped since the GDPR came into effect, and many registries put limits on how much data can be retrieved. All registries are required to provide a copy of their detailed registrant data to the Internet Corporation for Assigned Names and Numbers (ICANN) to assure continuity of domain name service in case the registry operator fails. However, ICANN does not make this data available for external analysis.

3.3 Security hygiene data

Various data can be gathered that reveal security aspects of various systems, or a view into the degree of compliance with recommended security practices by various actors in the ecosystem.

Blacklists (abuse data). Various providers generate lists (blacklists) of domains, URLs or IP addresses associated with abusive or malicious behavior, including spam, phishing and malware. Organizations that have shared such data with researchers include Spamhaus, Seclytics, and DomainTools. There are also public sources of such data, as described (and compared) in [24].

Source address validation data. An MIT graduate student – Robert Beverly, now at NPS – was sufficiently curious about academic assertions of pervasive SAV compliance that he developed a client/server software system to gather crowdsourced measurements, and personally maintained this infrastructure for years before deciding it needed an institutional home. With his assistance and U.S. research infrastructure funding, CAIDA hardened and operationalized this capability, enabling independent verification that a given network has properly deployed SAV. CAIDA used this platform to study the likely effects of deployed interventions to internalize this externality, e.g., naming and shaming. The unfortunate but unsurprising conclusion was that interventions tried thus far are not very effective at overcoming the strong counterincentives to compliance[25]. But any stronger interventions, including regulatory ones, will require this form of independent verification of SAV compliance.

4 ARTICULATING AN R&D AGENDA ENABLED BY AN OKN

We spent part of the workshop discussing specific research agenda questions, without reaching consensus on priorities. In the next workshop we will seek consensus on these questions, and drill down on what data is required to pursue them, and how to most effectively apply resulting knowledge for operational improvements.

4.1 R&D agenda for routing security

The workshop discussions made clear that there is no shortage of data about the routing system. But a knowledge gap persists. Despite 20 years of study, we still have no consensus on the prevalence and effectiveness of route hijacking attacks. There is no open knowledge of what ASes/prefixes, and types of ASes/prefixes, are being hijacked. A large obstacle is the challenge of discerning sophisticated traffic engineering from configuration mistakes from malicious intent. We also do not know how much harm hijacks realistically impose. Better measurement coverage is only one ingredient to tracking the frequency, reach, and impact of different types of hijacks over time. We also need to derive knowledge from the measurements, and use the knowledge strategically.

As an example of using measurements to advance routing security, we discussed an extension of the MANRS initiative (§3.1), which we tagged as MANRS+. MANRS+ would require empirical demonstration that a participating ISP in fact meets all of its commitments, and would require the ISP to commit to other security hygiene behavior, including dropping ROA-invalid announcements. MANRS+ would also disclose the failure of ISPs to conform to these norms. An attacker would thus not be able to use a simple hijack based on an invalid source-prefix announcement. They would have to use an invalid *path* announcement as a form of attack. In this form of hijack, a malicious AS announces an AS path with two (or more) hops, where it lists itself as the adjacent AS, and then falsely asserts that the origin AS is its customer (or multiple hops further away). The origin in this path looks legitimate, since it matches a registered ROA.

We discussed two possible mechanisms to deal with invalid path announcements. One is the recently proposed ASPA (Autonomous System Provider Authorization), in which customer ASes pre-register the ASes they will use as providers. These registrations allow an AS to detect if a path in a BGP route announcement is valid [2].

A second solution might be called *recursive MANRS*. If a MANRS+ compliant AS gets a BGP announcement containing more than one AS in the AS path announced, then either the AS sending the route announcement is MANRS+ compliant and has thus validated its own customer assertions, or the receiving AS must assume that the announcement may be bogus. In this case, if the receiving AS also has a route received from a MANRS+ compliant AS to the relevant prefix, it should just discard the potentially bogus assertion. If there is no competing route to the origin in the routing table of the MANRS+ compliant AS detecting the potentially bogus route, that AS could forward the route announcement but flag it as *dubious*, e.g., via a standardized BGP community value. Another AS receiving a dubious announcement could again forward it, but ASes should prefer paths that are not tagged dubious (similar to the idea proposed in [15]). This scheme depends on MANRS+

ISPs being directly connected with each other, in which case recursive application of this rule means that an attacker cannot succeed by creating a false path assertion.

This proposal triggers additional research questions. In principle, will a recursive-MANRS+ scheme work to prevent invalid path announcements? How much implementation effort would it require for code in current routers to support it? Can we create and perform a simulation or prototype that demonstrates this approach will work? What operational impediments could not be explored through simulation or prototype?

Since the scheme requires direct connectivity among MANRS+ participants, there are more immediate, but fortunately easier, questions: are MANRS members directly connected today? Can we analyze the topology around the known serial hijacker ASes [35]? Are they multi-homed (which makes path poisoning harder)? Are they served by the same provider ASes? What parts of the Internet are still susceptible to a particular hijack? A hijack can only propagate through a non-compliant region. What do these regions look like? Do different sorts of hijacks propagate differently? Would a denser deployment of BGP probes make the analysis more convincing? Is it possible to establish that MANRS+ compliance benefits the participating AS itself? Are customers of that AS less likely to suffer hijacks? Can other benefits be measured?

The overarching question posed is whether an sustained open knowledge network in this domain will catalyze the scientific advancement of data processing techniques to detect and classify different types of hijacks and misconfigurations, to facilitate manageability of individual networks, to enable verification that members of a collective such as MANRS+ behave properly, and to provide safeguards against corruption of the RPKI.

4.2 R&D agenda for naming security

The current state of the DNS ecosystem is dysfunctional, and escalating criticism of the current governance model could precipitate a crisis, including increasing pressure on ICANN to take action or risk its role in the world order. But any attempt to improve the security of the DNS system, and in particular to reduce its utility as a building block for abusive behavior, must consider options for abusers to strategically evolve in response to interventions. Most obviously, further descent into lawlessness and abuse will escalate of defensive action, e.g., aggressive blocking based on DNS or other criteria. Aggressive blocking could drive certain gTLD operators out of business. There is no consensus on the utility of different sorts of blocking or related controls, such as warnings, though potential collateral damage of DNS-based blocking is well-established [21]. Compounding this epistemological challenge, those who build blocklists (lists of domains or addresses to block) do not perform the blocking, and those who block do not normally discuss their decisions.

One response to aggressive blocking could be that registrars that serve spammers try to defend their rights–complain to ICANN or find someone to sue. More likely is that miscreants will find a new (or revert to an old) weak link, e.g., hijacking name servers and creating illegitimate domains. Miscreants already benefit from the complacency or complicity of registrars/registries, including by leveraging deeply discounted pricing and/or automated bulk registration. Registars that support and defend automatic provision of bulk registration of (e.g., thousands of) domains without verifying brand or other intellectual property compromises, is a well-establishd source of tension in the ICANN community. Owners often use these domains months later, after they have aged sufficiently to have somewhat trusted reputations. Escalated and more orchestrated blocking of domains to deal with abuse will bring attention to the operational aspects of DNS infrastructure, and likely lead to (further) concentration of the DNS infrastructure business onto a few large platform providers.

There is also the potential for deployment of some new component of the DNS architecture. Analogous to Google's successful push for certificate transparency, some organizations(s) might devise some mechanism for cross-checking or limiting abuse. Finally, and not mutually exclusive with any other options, miscreants might avoid using the DNS, and use URLs with embedded IP addresses. In fact, any app may exercise control over name resolution, e.g., use a custom resolver, or avoid the DNS entirely.

As with routing security, this discussion emphasized the lack of any quantitative baseline description of many aspects of the DNS, which makes it difficult to show that circumstances are deteriorating. Can we conceptualize, and construct an annotated map of the namespace, or a subset of interest, including zone creation, expiration and configuration patterns that may represent security or resilience vulnerabilities [4]? Can we create an open, aggregate, anonymous reporting system that indicates how many recursive resolvers are blocking which gTLDs? A related challenge is to map out the money flows in the DNS ecosystem, and the contribution of bad actors to the flows.

Measurement of harm is even more challenging. Can we measure or predict collateral damage of blocking, using logs of queries to recursive resolvers? This would require methods to discern which queries to suspicious gTLDs are to malicious sites, and which are legitimate. Can one reduce collateral damage by making blocking decision on a regional basis? What would global providers like Google do, assuming that they implement a blocking regime?

What practices by responsible registries can help? Should registries be allowed to choose which registrars they use for their business? (This would require changes to ICANN bylaws.) Can bulk registrants be tracked through credit card number (or a one-way hash thereof, to protect privacy)?

Finally, can we map out domains of power in resolution? Traditionally, ISPs (by means of DHCP) have controlled the choice of resolver, although savvy users can configure their OS to use a different resolver. This choice matters because recursive server operators have the power to block domains, or to use an alternative root. The recent DNS-over-HTTP debate has made it clearer that various actors can control which recursive resolver is used, in particular the browser provider, and by the same argument, any native app in a mobile device, which is where the future is heading. How could defenders counter a DNS-bypass approach by miscreants?

Many organizations and groups are actively considering Internet identifier security issues [11, 12, 28]. One goal of this workshop was to identify a range of data sets that could serve as the basis of an Open Knowledge Network, and the range of ways that existing players might participate in such an OKN, as imagined in NSF's Convergence Accelerator Program [29, 30].

5 ORGANIZATIONAL SUPPORT

A challenge to developing and maintaining an OKN is the need for sustainability for its data components. We identified three institutional models for organizations that collect and curate data to support security research on these systems.

Governments. Similar to other critical infrastructures (energy, transportation, water, food, finance), governments may collect data, or fund collection of data related to security and resilience. One example of this in the U.S. is NIST's RPKI deployment monitor.[2]

Non profits and academics. Academic researchers sometimes curate and share Internet measurement data. History has shown such arrangements to be fragile. One person may drive them, perhaps a PhD student, and when that student leaves, the infrastructure may be difficult to maintain. The Spoofer project, DNSCoffee, RouteViews are all examples of this sort of project. If maintained, such projects require continual fund-raising, and are often underfunded which constrains what they can do.

OARC is supported through corporate membership, with modest but stable funding that supports the DITL collection (§3.2). However, OARC members tend to prioritize its open source software efforts and workshops more than the DITL data collection, so its future in uncertain. Similarly, RIPE's data collection project are funded by RIPE's membership fees, and face membership pressures and constraints. Their focus is not supporting scientific research although they are supportive of this use of their data.

ICANN is a special organization, holding responsibility for stewardship of the top-level of the Domain Name System, and overall stewardship of DNS governance. ICANN is supported largely through the share of revenues it receives from the sale of domain names (and to a lesser extent, the registration of IP addresses). Two recent reviews have criticized ICANN for not enabling sufficient transparency with data that it has or could easily obtain [1, 33].

Commercial firms. The persistent insecurity of the Internet has led to a large ecosystem of firms that gather data from which they derive and sell *threat intelligence* data feeds to customers, who use them to configure security services. An example represented at the workshop, Farsight is a for-profit threat intelligence firm, where the management supports research use of the data as much as is feasible. Other commercial organizations collecting such data include Spamhaus, Secalytics and Cisco Umbrella.

There are three issues with data collected by commercial firms. First, the pricing models may prohibit use by scientific researchers. Second, the data is typically organized to answer specific queries, related to real-time detection of threats and forensic analysis of recent incidents, which makes it difficult or impossible to use for other purposes, including longitudinal trends. For example, a commercial provider of data may make a free version available, based on limiting the number of queries per day. Analyzing long term trends requires a large corpus of historical data, which is not possible with this access model. Some firms do make data available for non-commercial research, but each firm has its own data usage restrictions. Working with multiple data sets may require complex negotiation, and may result in severely restrictive usage rules. Third, commercial data in a third-party's hands may reveal something that poses a threat to the commercial owner of the data, a counterincentive to sharing it.

Additionally, data collection and sharing operate in an environment characterized by increasing concerns about privacy, as manifested in the European General Data Protection Regulation (GDPR) or the recent California Consumer Privacy Act (CCPA). These laws add difficulty and uncertainty to data collection and sharing. IP addresses are usually considered PII and many datasets contain IP addresses, which inhibits (if not prohibits) sharing them. One unintended consequence of GDPR was a still-unresolved clash for DNS registrars that sell domain names. Their contracts with ICANN require disclosure of registrant contact information (e.g., address) but one interpretation of GDPR suggests that such disclosure requires an appropriate legal order. As a result, registrars who may serve European registrants (i.e., most registrars) have decided to withhold any potentially personal information about domain name owners. This decision has rendered analysis of domain abuse by third-party researchers more difficult. The vetting required to confirm a trusted use of the data can be expensive, creating a revenue opportunity for registrars (if they could charge the trusted user), but also upsetting a delicate balance in the security research ecosystem, giving attackers a decided advantage.

6 NEXT STEPS

This workshop established common ground regarding the range of data available, how it is used and supported, and understanding of the chasms between raw data, appropriately curated data, and scientific knowledge related to Internet infrastructure security and stability. We focused on threats and intelligence about traditional core infrastructure of the Internet (BGP, DNS, IP addresses), but the Internet's evolution suggests a diminished role for these core protocols. Direct peering between content and access giants means that most Internet traffic is along BGP paths of length 1. The evolving model of the DNS (DOH/ABCD) is such that resolution paths may follow a similar model. What are the implications for scientific study of the infrastructure?

We also need to consider new models of Internet usage: smart cities, fog computing, 5G, etc. What usage patterns, control planes will be important to an internet where the majority of communication is M2M or M2 edge compute node? What is the role of DNS and BGP critical infrastructures to that emerging world?

The goal for the second workshop (February 2020) will be to clarify the research agenda we hope KISMET will enable, and the data sets, analytic capabilities, and organizational needs to achieve it. We will use hypothetical future scenarios to motivate research questions, and try to identify how to measure that specific scenarios are occurring, or how to measure resulting harm.

Finally, a knowledge network is more than a collection of data sets. We will need to identify metadata that can enable discoveries that cut across data sets. Given the explosion of interest in machine learning, we will consider how to position a knowledge network to serve this emerging trend. We hope this strategy will allow the security community to go beyond consideration of attacks that have already been seen in the wild, since it is clear that vulnerabilities could be exploited in ways that we have not yet seen.

[2] https://rpki-monitor.antd.nist.gov/

7 ACKNOWLEDGMENTS

We are grateful to the participants for contributing their insights at the workshop, and for feedback on this report. This workshop was supported by NSF C-ACCEL OIA-1937165. Opinions expressed do not necessarily reflect views of the NSF. Any errors are the responsibility of the authors. We will publish a final report that summarizes the output of both workshops in April 2020.

REFERENCES

[1] Accountability and Transparency Review Team 3. Third Accountability and Transparency Review Team (ATRT3) Draft Report, 11 2019. https://www.icann.org/en/system/files/files/draft-report-atrt3-16dec19-en.pdf.
[2] A. Azimov, E. Uskov, R. Bush, K. Patel, J. Snijders, and R. Housley. A Profile for Autonomous System Provider Authorization, Nov. 2019.
[3] F. Baker and P. Savola. Ingress filtering for multihomed networks, Mar. 2004. IETF BCP84.
[4] S. Bates, J. Bowers, S. Greenstein, J. Weinstock, and J. Zittrain. Evidence of Decreasing Internet Entropy: The Lack of Redundancy in DNS Resolution by Major Websites and Services. Technical report, Harvard University, 2018. https://dash.harvard.edu/bitstream/handle/1/35979525/DNS_NBER_Working_Paper.pdf.
[5] Brian Cute. Evolving the effectiveness of our multistakeholdermodel, Mar. 2019.
[6] Center for Applied Internet Data Analysis. CAIDA Internet Data. http://www.caida.org/data/.
[7] Center for Applied Internet Data Analysis (CAIDA). Prefix to AS mappings. http://www.caida.org/data/routing/routeviews-prefix2as.xml.
[8] T. Chung, E. Aben, T. Bruijnzeels, B. Chandrasekaran, D. Choffnes, D. Levin, B. M. Maggs, A. Mislove, R. v. Rijswijk-Deij, J. Rula, and N. Sullivan. RPKI is Coming of Age: A Longitudinal Study of RPKI Deployment and Invalid Route Origins. In *IMC*, IMC '19, 2019.
[9] Cisco. BGPMON. http://bgpmon.net/, 2018.
[10] D. Clark and k. claffy. Toward a Theory of Harms in the Internet Ecosystem. In *Telecommunications Policy Research Conference (TPRC)*, Aug 2019.
[11] Competition, Consumer Trust, and Consumer Choice Review Team. Final Report, 2019. https://www.icann.org/en/system/files/files/cct-rt-final-08sep18-en.pdf.
[12] Dave Piscitello and Dr. Colin Strutt. Criminal Abuse of Domain Names Bulk Registration and Contact Information Access. http://www.interisle.net/sub/CriminalDomainAbuse.pdf.
[13] P. Ferguson and D. Senie. Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing, May 2000. IETF BCP38.
[14] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos. A Look at Router Geolocation in Public and Commercial Databases. In *ACM Internet Measurement Conference (IMC)*, Nov 2017.
[15] P. Gill, M. Schapira, and S. Goldberg. Let the Market Drive Deployment: A Strategy for Transitioning to BGP Security. *SIGCOMM Comput. Commun. Rev.*, Aug. 2011.
[16] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker. From.academy to.zone: An Analysis of the New TLD Land Rush. In *Internet Measurement Conference (IMC)*, 2015.
[17] T. Halvorson, K. Levchenko, S. Savage, and G. M. Voelker. XXXtortion?: inferring registration intent in the .XXX TLD. In *Internation Conference on World Wide Web*, Apr. 2014.
[18] B. Huffaker, K. Keys, M. Fomenkov, and K. Claffy. AS-to-Organization Dataset. http://www.caida.org/research/topology/as2org.
[19] Ian Foster. dns.coffee. https://dns.coffee/.
[20] ICANN. Expedited Policy Development Process on the Temporary Specification for gTLD Registration Data, 2018.
[21] ICANN Security and Stability Advisory Committee. SSAC Advisory on Impacts of Content Blocking via the Domain Name System, Oct 2012. https://www.icann.org/en/system/files/files/sac-056-en.pdf.
[22] ICANN Technical Study Group, led by Ram Mohan (Afilias). TSG001: Draft Technical Model for Access to Non-Public Registration Data, 3 2019. https://www.icann.org/en/system/files/files/draft-technical-model-access-non-public-registration-data-06mar19-en.pdf.
[23] M. Kührer, T. Hupperich, C. Rossow, and T. Holz. Hell of a handshake: Abusing TCP for reflective amplification ddos attacks. In *8th USENIX Workshop on Offensive Technologies (WOOT 14)*, San Diego, CA, 2014. USENIX Association.
[24] 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 *USENIX Security*, 2019.
[25] 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.
[26] 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.
[27] S. Lyngaas. Someone is spoofing big bank IP addresses – possibly to embarrass security vendors. *Cyberscoop.com*, Apr. 2019.
[28] Messaging, Malware and Mobile Anti-Abuse Working Group. AAWG Tutorial on Third Party Recursive Resolvers and Encrypting DNS Stub Resolver-to-Recursive Resolver Traffic, Version 1.0, 9 2019.
[29] National Science Foundataion. Convergence Accelerator Program, 2019. https://www.nsf.gov/od/oia/convergence-accelerator/index.jsp.
[30] Networking and Information Technology Research and Development Program. Open Knowledge Network: Summary of the Big Data IWG Workshop, 2017. https://www.nitrd.gov/nitrdgroups/index.php?title=Open_Knowledge_Network.
[31] nist.gov. Nist rpki monitor. https://rpki-monitor.antd.nist.gov/.
[32] Paul Vixie. Rate-limiting State: The edge of the Internet is an unruly place. *ACM Queue*, 12, February 2014.
[33] Security and Stability Review Team. Draft recommendations. https://66.schedule.icann.org/meetings/1116775.
[34] K. Sriram, D. Montgomery, D. R. McPherson, E. Osterweil, and B. Dickson. Problem Definition and Classification of BGP Route Leaks. RFC 7908, June 2016.
[35] 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.
[36] U.S. Dept. of Homeland Security. Emergency Directive 19-01: Mitigate DNS Infrastructure Tampering, Jan. 2019.
[37] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras. A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurements. *IEEE Journal on Selected Areas in Communications*, 34(7), 2016.