Integrated Library for Advancing Network Data Science - (ILANDS)

1 Rationale and Need for Infrastructure

Understanding the Internet's changing character is impossible without realistic and representative datasets and measurement infrastructure that can support sustained longitudinal measurements as well as new experiments, and with resulting data available to scientific researchers. But there is a dearth of good data to support research, for several good reasons: complexity, scale, and cost of measurement instrumentation; information-hiding properties of the routing system; **privacy**, security and commercial sensitivities; costs of storing and processing the data; and lack of incentives to gather data in the first place, including cost-effective ways to use it operationally [1]. This opacity of the Internet infrastructure hinders research and development efforts to model network behavior and topology, and design protocols and new architectures. More fundamentally, it also hinders our ability to understand and reason about real-world properties of the Internet such as robustness, resilience, security, and stability.

We propose to upgrade and integrate two of our measurement capabilities to enable a community of researchers across many institutions to collaborate on a high-level focused agenda: **understanding the evolving character of the Internet, with the objective of assessing, modeling and shaping its security and resilience**. This infrastructure will also enable CAIDA and RouteViews to continue their 20+ year record of providing high-quality measurement data to enrich the scientific network research endeavors of a diverse community of investigators.

Our first measurement capability is traffic capture. For decades it has been virtually impossible for researchers to get access to passive (sniffed) **traffic data** from Internet backbone links due to privacy concerns. Based on trust relationships that have been maintained for over two decades, CAIDA has been able to measure strategic links in the backbone so long is it could provide funding for the monitor. We had to stop these data sets in January 2019 when the link upgrade to 100GB rendered collection technology obsolete. The second measurement capability relates to the **structure and dynamics of the network topology**. The RouteViews project has been sharing Internet routing data with the research community for more than twenty years – while the Internet routing table has increased an order of magnitude in size. We propose to enhance infrastructure to sustain the increased packet rates and routing table growth, expand storage and compute resources to support long-term use of the data, and develop additional tools to facilitate their combined use. While these two data sources have existed separately in the past, current privacy-protection approaches have prohibited joint analysis across them. A concomitant goal of the proposed effort is to investigate approaches to protect privacy that will permit integration of the gathered data, a new capability that is critical to some of the identified research goals.

Few research teams would be able to field this infrastructure much less support sharing this data. A collaborative team at CAIDA and NSRC will develop user services and engagement needed to grow a robust research community that is actively involved in determining directions for the infrastructure, and optimizing NSF's investment toward achievement of these goals. Our outreach coordination process will have five objectives: (1) shape what data we collect and store, its formats, access model, long-term archival, and overall management of the infrastructure; (2) find new users of the infrastructure, especially from underrepresented groups, and assist them with sound use of the data; and (3) bring this community together to facilitate collaboration on a focused research area that would not otherwise be possible (Section 2); (4) publish methodologies of data analysis and resulting insights; (5) establish a concrete plan for sustainability of the infrastructure to support the needs of the U.S. science and engineering research community.

2 Focused Research Agenda: Understanding the Evolution of the Internet

Consistent with the goals of the CCRI, we propose a bold research direction, which this infrastructure will facilitate and catalyze: **understanding the evolving character of the Internet, with the objective of assessing, modeling and shaping its security and resilience**. This challenge is long-standing and daunting.

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. Society needs a more rigorous understanding of the Internet ecosystem, a need made more urgent by the rising influence of adversarial actors.

For systems critical to society, such as health care, transportation, agriculture, and commerce, 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. 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.

The daunting challenge is that understanding of the structure and dynamics of Internet topology, routing, workload, performance, and vulnerabilities require large-scale distributed network measurement infrastructure. A CCRI project will not completely solve this challenge, but we can make significant advances, and catalyze Internet research that would be difficult or impossible without the proposed infrastructure. In particular, the Internet research community can benefit greatly from greater knowledge and access to advanced methods of data science, which will require data sets to experiment with them.

We propose to build instrumentation that can enable a focused group of researchers to provide empirically grounded answers to critical questions, including questions that can lead to measurable improvements in security and resilience of the Internet infrastructure.

"Regionalization" of Internet traffic and connectivity. The design goal of the Internet was and continues to be that any two machines anywhere on the Internet could freely communicate. A packet might cross several ASes to reach its destination, but today most traffic traverses only one, in large part due to the goal of efficient delivery of high-volume content from large providers, e.g., Netflix, Amazon, and YouTube, to access providers. Not only does the traffic need to cross only one service provider, but extensive use of caching means that traffic enters the access network at a point close to where it will exit to the consumer. Understanding of this trend is critical to the development of new evidence-based approaches to improve the security of the Internet infrastructure. Measuring the degree to which the Internet experience has become more localized is challenging, because this property depends acutely on where that user is within the Internet. A user attached to a large U.S. broadband access provider will probably have a more localized experience than a

user from the developing world.

Routing security. At any moment, there are about 850k routing assertions being injected into the Internet. Security researchers use routing data such as that collected by RouteViews to attempt to identify the few of them that may be malicious. But the challenge is more than finding a few bad needles in a haystack of 850k needles, it requires finding malicious *changes* in a constantly changing haystack, which also requires understanding *baseline* behavior for a given time and network. To make the challenge harder, different measurement points across the Internet will pick up different variants of these assertions, because as the assertions propagate through the Internet, every router will modify them, and pick a subset of them to store and propagate, according to local rules that are not visible. External measurement can observe consequences of these local decisions, but does not have access to the rules themselves. Furthermore, since normal practices evolve rapidly on the Internet, any model of baseline dynamics quickly ages.

Toward new approaches to thinking about security: zones of trust. Past attempts to remediate Internet routing protocol vulnerabilities have considered technical remedies, such as protocol enhancements. A purely technical approach has proved unsuccessful. First, the global and multistakeholder nature of protocol development makes consensus difficult or impossible. More problematic, proposing or even standardizing a new technology does not mean that actors will deploy it. This reality implies that the path to better security does not lie in proposals for global changes to the Internet protocols, but in finding *operational practices* that *regions* of the Internet can implement to improve the security profile of those regions. This approach allows groups that choose to trust each other to define and circumscribe the systems they trust. It is more consistent with trust models in the physical world, where we accept that there are malicious actors, and we attempt to arrange circumstances to minimize our interaction with them, and to interact with potentially untrustworthy actors only in constrained ways. Clark and Claffy have recently referred to regions that embody a common sense of commitment to distance themselves from the global pool of bad actors as *zones of trust* [2]. To be effective, this approach must include mechanisms to keep typical activities of users inside such a zone. The success of such an approach depends in part on how topological characteristics of the Internet are evolving, and will also require ongoing measurements of networks in the zones.

Discovery of botnet location, scope, and prevalence. Botnets play a critical role in adversary reconnaissance (scanning and phishing), influence operations (up voting), and financing (ransom ware, market manipulation, denial of service, spamming, ad clicks) [3]. The advent of the Internet of Things amplifies the scope and potential damage due to botnets. The application of advanced analytics to the the complex intellectual challenge of finding the traffic signature of a botnet in traces of billions of packets has shown promising early results using CAIDA's previous traffic data [4]. This data has allowed researchers to demonstrate that "significant progress can be made using AI approaches that operate on anonymous data, leading to a potential billion-fold reduction in the amount of deanonymization required [4]."

Network tomography to optimize traffic monitor deployment. Traffic data can be used for a variety of purposes, as we list above. A challenge for the deployment of traffic monitors proposed here is where best to put them to capture the most useful data, which may vary by objective. Deployment will be shaped by many practical factors, including the technical complexity and business considerations associated with a particular location. While we develop the collection

platform, we will use routing (RouteViews) and topology (CAIDA traceroute) data to identify a set of locations that are diverse in terms of the sorts of traffic that we expect to see passing that point. With a good understanding of network topology, we believe that we can identify points that will be "typical", in that the traffic captured there will resemble traffic that might pass by many points in the Internet.

Macroscope measures of cyberspace Beyond the specific questions about security and resilience, this infrastructure can help inform more general insights about the character of the Internet, which are increasingly relevant as society becomes ever more dependent on this infrastructure: how big is the Internet, by various measures? how is it growing? how are global and regional traffic patterns changing? how is the mix of applications changing? what can we learn about the character of the Internet in different parts of the world?

Validation Researchers and network engineers develop abstract models of the Internet at various levels of detail, which are then used to support large-scale simulations, development of new protocols and architectural concepts, and economic analyses. Data from this infrastructure can help confirm the validity of those models, and improve the foundations from which that subsequent research grows.

2.1 Community involvement in research agenda

The letters of collaboration demonstrate substantial involvement of CISE researchers to advance our focused research agenda. In addition, a diverse community of investigators will find the proposed infrastructure valuable to their individual research endeavors. Table 1 lists a sample of research interests articulated in the attached letters of collaboration, and which proposed infrastructure enhancements are required to support them.

3 Infrastructure Description

Our infrastructure enhancement has three components: capture, management, and sharing of Internet traffic header data; capture, management, and sharing of interdomain BGP routing data; and tools and user services to lower the barrier to use of this data.

3.1 Existing infrastructure

Infrastructure to Gather and Share Traffic Data. Since April 2008, CAIDA's passive traces dataset contains traces collected from high-speed monitors on a commercial backbone link, and anonymized for sharing with the research community [5]. Six times in the last 20 years years this backbone infrastructure has upgraded beyond the scope of the budget CAIDA has for monitoring (OC3, OC12, OC48, and OC192, 40GB, and now 100GB). As of January 2019 CAIDA has not been able to capture data on Internet backbone links due to lack of resources; the link was upgraded to 100GB and we did not have funds to upgrade the monitor. CAIDA's (and, to our knowledge, the Internet's) last remaining single last point of public insight into the commercial Internet backbone was lost in January 2019.

The traces are anonymized with CryptoPan prefix-preserving anonymization [6]. The anonymization key changes annually and is the same for all traces recorded during the same calendar year. During capture, packets are truncated at a snap length selected to avoid excessive packet loss due to disk I/O overload. The snap length has historically varied from 64 to 96 bytes. Payload

Table 1: Focused research agenda topics described in letters of collaboration (LOCs). Undergraduate-only institutional collaborators include Sommers (Colgate) and Arnold (West Point).

Enabled Research	**Collaborators**
Understanding Internet Security Vulnerabilities	
· IoT device security	Caesar
· Botnets	Kepner
· Routing Security	Testart, Wahlisch, Schmidt, Freedman, Camp
· Transport Security	Camp
· Role of DNS	Allman
· DDoS attacks	Wahlisch, Schmidt, Sollins
· Anomaly detection	Springer, Ramakrishnan, Bishop
· Traffic characterization and classification	Berger, Hussain, Bishop
· Blacklist validation	Levchenko
Understanding Internet Stability and Resilience Properties	
· Evolution of congestion control protocols	Steenkist, Sheery
· Evolution of traffic characteristics	Tseng
Internet Topology Structure & Dynamics	
· Evolution of Internet topology	Hussain, Carisimo, Arnold, Katz-Bassett
· Regionalization of traffic	Clark, Latour, Allman, Arnold
· New tools to support Internet Topology Inference	Luckie
Infrastructure Architecture & Evolution	
· Outreach	Clark, Huter
· New approaches to sustainable measurement infrastructure	Kentik, Levchenko, Zhang, Rexford
· MetaData Generation	Aben, Sommers

is removed from all packets. We have used Endace network cards to record these traces, which provide timestamps with nanosecond precision. However, the anonymized traces are stored in pcap format with timestamps truncated to microseconds. Starting with the 2010 traces the original nanosecond timestamps are provided as separate ascii files alongside the pcap files. The traces can be read with any software that reads the pcap format, including the CoralReef Software Suite, tcpdump, Wireshark, and many others. These data are useful for research on the characteristics of Internet traffic, including application breakdown, security events, geographic and topological distribution, flow volume and duration. It also supports development of new traffic analysis technologies.

Infrastructure to Gather and Share Routing Data. The Internet's global routing mechanism (the Border Gateway Protocol or BGP) uses *routing assertions*, created by the seventy thousand regions (*autonomous systems*) of the Internet to describe the Internet (IPv4 and IPv6) addresses in each region. The size of the IPv4 routing table has more than tripled in the past decade from 250,000 entries in 2009 to 840,000 entries today. The global IPv6 routing table entries increased from 1,100 entries in 2009 to nearly 100,000 IPv6 routing entries today. RouteViews started its first BGP routing data collection in 1995 and has continuously archived collected data since 1997, which contains the oldest data sets in existence.

Although originally intended for operational rather than research use, the RouteViews data has become indispensable for an astonishing range of Internet research. However, it has always survived on soft-money project funding and occasional donations of equipment from industry over last two years. This longest-standing BGP data collection infrastructure is at an inflection point, having had its operating personnel retire two years ago, and has been limping along for years without upgrades or expansions despite the Internet's relentless topological growth.

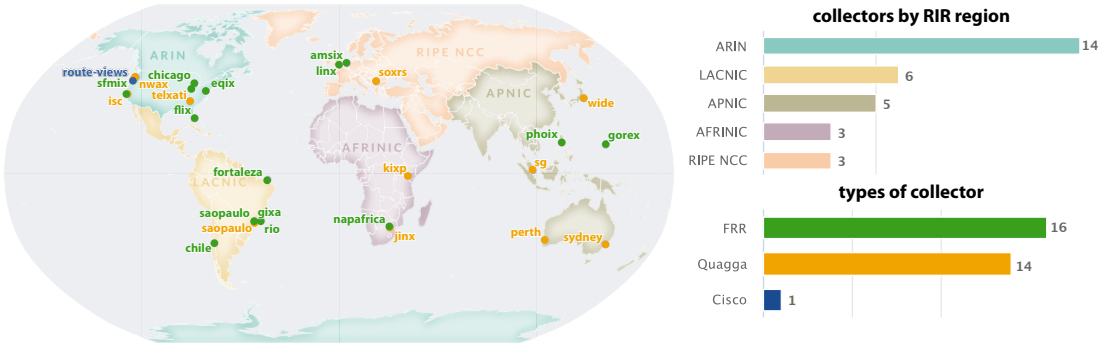

Figure 1: Routeview collector locations include London, Sydney, Singapore, So Paulo, Santiago, Nairobi, South Africa, plus Northern Virginia and Palo Alto, California in the U.S. Between them, they collect more than 600 active routing feeds. The infrastructure has true global coverage.

3.2 Plan for enhancement/sustainment of the infrastructure

3.2.1 Enhancements to Gather and Share Traffic Data

CAIDA proposes to build at least two 100GB passive monitoring platforms to capture data to support the focused research agenda (Section 2). We already have agreements in place to allow access to strategic infrastructural locations in backbone providers, tools for anonymizing the data, and are now only blocked on funding for the monitors themselves, storage, and personnel to build and maintain them and manage the data.

Hardware. We plan to use two servers: the Capture machine, and the Compute machine. The capture machine hosts two 100G cards. The first is the actual capture card (Napatech NT200A02) and a transfer card (Mellanox ConnectX-5). The Compute server has one 100G transfer card (another Mellanox ConnectX-5) and 8 1.6TB SSDs used to store actual captures. Traffic will be time-stamped with nanosecond precision. The capture machine will take data gathered from the capture card, strip payload, and send it to the compute server by way of the Mellanox card. The compute server runs 8 or more receiver processes that take incoming packets from the capture server and stores them on one of the 8 SSDs, as well as taking various Metadata about these traces and uploading them to our time series databases.

Data access. We will build a system that indexes flow-level data for interactive querying. We will leverage the virtualized cloud research compute environment we built for the UCSD Network Telescope [7] to support researchers who cannot download traces for local use.

Data privacy In the existing infrastructure, CAIDA has addressed privacy concerns by using the CryptoPan scheme to anonymize IP addresses. This scheme has been accepted as providing sufficient privacy protection; however it does not permit certain sorts of queries, such as exploration of the regionalization of traffic patterns over time. One option for the new packet capture infrastructure is to use different schemes for privacy protection at different times, thus producing disjoint data sets useful for different experiments. Our research collaborators will play an early role in the design of the collectors; together we will analyze different schemes for privacy protection that are acceptable to the providers of the raw packet feeds and as well useful for different sorts of experiments.

Partnerships and sustainability. CAIDA will collaborate with Kentik [8] to use their commercial traffic monitoring platform to augment our own infrastructure, and possibly as a way

to reduce the amount of infrastructure that CAIDA needs to operate on an ongoing basis. Kentik provides a cloud-based platform that captures and indexes full-resolution network data and provides an interactive, web-based query and visualization interface. Kentik has provided free access to their platform for UCSD's Network Telescope data [7], and we used it to identify several networks responsible for a DDoS attack and block them at the UCSD campus border. Kentik is interested in facilitating use of their aggregated traffic data stream of about 10million flows/sec from 300 ASNs to answer questions about the current aggregate characteristics of Internet traffic. This heavily aggregated data will not be sufficient for many of the questions in the research agenda, but it potentially provides an opportunity for validation of whether our data sets are representative of the larger Internet.

3.2.2 Enhancements to Gather and Share BGP Routing Data

Hardware. We propose to add new collectors in locations of interest to the research community. The milestone set for each year will be to deploy between three and five new collector locations in new and different countries/regions of the world. Additional funding is budgeted to replace some older collectors with new hardware, which is necessary to fully utilize new software features. When possible, RouteViews utilizes virtual machines (VMs) at host locations to conserve funding.

Enhanced Infrastructure Capabilities. We have separately proposed to leverage existing BGP collector infrastructure to support active network measurements, using CAIDA's Ark measurement software. This capability allows for correlation of data plane and control plane views from the same vantage point, which has been a long-standing visibility gap in the Internet research community. If that separate project is funded, this software will be fully integrated across the global RouteViews software stack by the end of Year 1 of this project. But this software will require a user interface to enable community access to the capability. We propose as part of this project to create this user interface based on a previous software effort to create a unified interface to looking glass servers [9]. Researchers will then be able to monitor, analyze, and remotely trigger active measurements from a common set of tools and applications, resulting in new opportunities for data science and machine learning techniques.

Enhanced Availability. CAIDA operates a service broker which assists BGPStream applications in identifying and acquiring real-time and historical BGP data. The service broker presents a single point of failure in the operation of BGPStream libraries and analysis applications. To address this important infrastructure component, RouteViews will deploy and operate an additional service broker to eliminate the current single point of failure. In collaboration with CAIDA, we propose to modernize the service broker architecture and code base, to make it modular and extensible. This will allow the research community to extend the service broker to meet future research needs, such as inclusion of additional BGP data feeds, and increase reliability of BGP monitoring infrastructure. We will also enhance CAIDAs libbgpstream library, which works in conjunction with the BGPStream service broker to facilitate analysis of historical and real-time BGP data, and to integrate complementary data sources such as RPKI [10].

Access to Enhanced Data Products. The RouteViews research platform is currently being augmented to utilize the IETF-standardized BGP Monitoring Protocol (BMP) software and provide an API for access to the collected data. With enhancements to route collector software and expansion of the database architecture, implementing BMP results in a new interface to stream telemetry from the global routing system and an API for consuming a steadily growing dataset of routing information. These new capabilities will be operational in Year 1 and provide significant benefit to both the operational and research communities trying to better understand global routing dynamics.

Partnerships. We are in discussion with Regional Internet Registry personnel at the RIPE NCC in Europe who manage RIPEs Routing Information System (RIS) collectors to better integrate collector data and access for the U.S. research community. Resilience and redundancy of data sources shared between RouteViews and RIPE RIS enhances verification and analytical capabilities of BGP routing data.

Sustainability. NSRC will lead efforts to develop a more sustainable model for the BGP monitoring and routing security cyberinfrastructure, including partnerships with industry groups that use RouteViews infrastructure and data, such as Google, Amazon, Microsoft and others to provide high leverage of NSF funds.

3.3 Tools, resources, and data sets

We will enhance existing and develop new supporting resources to integrate into the infrastructure, to facilitate optimal use of the infrastructure and enhance its value to the community. Our plan is to run a series of capture experiments, each gathering and aggregating data in different ways that protect privacy while allowing a range of experiments. For our raw packet capture experiment, we will aim to gather one-hour traces once per quarter. For this capture mode we will continue to use current tools for anonymizing the traffic data (Section 3 .1). However, a key limitation of this anonymity scheme is that it does not allow us to correlate traffic and routing data, because of the way IP addresses are anonymized. We will work with the research collaborators to devise additional ways to aggregate and anonymize packet traces so that with certain traces investigators can perform integrated analysis of traffic and routing data.

Tool enhancements to support enclave use. CAIDA will adapt Corsaro FlowTuple software to support two-way traffic, and adapt it to support publication of two-way flows to a Kafka cluster. This plugin will allow users to process a traffic sample as a stream and consume the derived flow data as soon as it is available, and run the traffic data through their own analysis scripts. (Section 3.2.2).

3.4 User services

The most important user service that CAIDA has provided for twenty years are our data curation and sharing processes. CAIDA invests significant effort to ensure that petabytes of data that we gather or derive each year are available to other researchers. We list all our datasets in our "Data Overview" table [11] and in the CAIDA catalog [12]. For each dataset we maintain metadata and all other relevant information including webpages, user guides, processing scripts, and previous publications. CAIDA datasets fall into two categories: public and by-request. Public datasets are available to users who agree to CAIDA's Acceptable Use Policy for public data. The by-request datasets are available for use by academic researchers, U.S. government agencies, and corporate entities who participate in CAIDA's membership program. Users fill out the request form and provide a brief description of their intended use of the data, and agree to an Acceptable Use Policy. We carefully review all requests to make sure that researchers understand the data they are requesting, and the use is in accordance with CAIDA and UCSD policies. We are now sharing some of these data sets via a Globus-authenticated [13] Swift storage cluster at UCSD.

Supporting analysis of data with SDSC and UO local cloud resources. At 100GB rates at typical link utilization, a one-hour trace of packet headers requires between 1-2 Terabytes of disk space. Many users will not be able to download such large data sets. We will use the tools described in

Section 3.3 to curate our traffic data, and support a Kafka cluster and virtual machines local to SDSC and UO to access the data.

Index data sets in CAIDA's data catalog. A major problem for researchers is the ability to find and understand available datasets. CAIDA's data resource catalog [12] addresses both of these problems by providing a unified, annotated, and searchable catalog of datasets, and a graph containing the relationship between these objects and dataset specific recipes for understanding, combining, and using CAIDA data. We will expand the catalog with the addition of publications that use products of the proposed infrastructure, and other datasets used in those publications. We will augment the current manually curated metadata with the option for programmatic updates via a web API. This will allow third parties to update metadata automatically.

We will also explore the possibility of using machine learning to process the thousands of publications that exist in our field. We will examine the degree to which we can apply existing research [14] to Internet-related publications and datasets. This enhancement would allow us to expand the catalog's scope beyond what we can support manually.

3.5 Community engagement

We have structured this project to provide the research community a continuing voice in the future directions, organization, and management of this infrastructure. We are committed to publicly releasing our analysis, visualization, and query tools. As required by the solicitation, we will participate in the CCRI Virtual Organization and community PI meetings.

1. **Ongoing user support.** A Data Adminstrator will handle all data inquiries and requests, and also review all proposed uses of data to make sure the use matches the data requested. Our Data Administrator generally responds to data requests within 48 hours, and regularly responds to questions sent to data-info@caida.org. CAIDA maintains mailing lists for researchers who have requested our data as well as a public list for general announcements regarding CAIDA data. We will use Mattermost (an open-source Slack equivalent) to support real-time communication with users.
2. **Biannnual newsletters and meetings.** Twice a year we will host *virtual community meetings* to support users or potential users of the ILANDS infrastructure and data. The Data Administrator will publish newsletters to the community in advance of these meetings, and send them to the community mailing list. These newsletters will include quantitative metrics of progress against community goals: data sets, data users, publications, cross-community efforts, and relevant use cases.
3. **Annual workshops and survey.** We will organize, lead, and write reports for **annual community workshops** to discuss measurement and querying capabilities, data formats, aggregation methods and other data curation functionality.
   Per our usage agreements for each protected dataset, we conduct periodic (at least annual) **surveys of our data users** to request a summary of research results. We also solicit feedback on the usability of specific datasets, difficulties users had with the data, and what new datasets researchers would like.

3.6 Community Outreach

Our outreach coordination process is designed to not only find new users of our data, but bring this community together to share results from using the data to support the focused research agenda, as well as methods of using it.

The Outreach Coordinator will lead efforts to build bridges and synergy with other communities other disciplines, via presentation of ILANDS resources at other workshops and conferences. An understanding of topology, routing and traffic is relevant to disciplines ranging from economics to political science, but researchers in those fields are not normally equipped with the skills to do their own analysis of the data we capture here. In most cases, scholars in other fields are best served either by careful translation of research results into a form that is useful to them, or by forging a collaboration among researchers from multiple disciplines. This sort of matchmaking is becoming more central to making progress in areas that touch society as the Internet does, and will be a priority for our outreach program.

We will also need to establish a tighter relationship with the data privacy technology research community e.g., multi-party computation, differential privacy. While many papers discuss these mechanisms, they have limited use by the practicing Internet measurement community. A goal of the outreach program is to understand why, and perhaps create more effective ways to integrate these tools into the practice of network measurement.

Finally, we will engage industry, intially encouraging them to establish BGP collectors and brokers, and to connect researchers with industry sources of traffic data, using ILANDS data protection tools and methods.

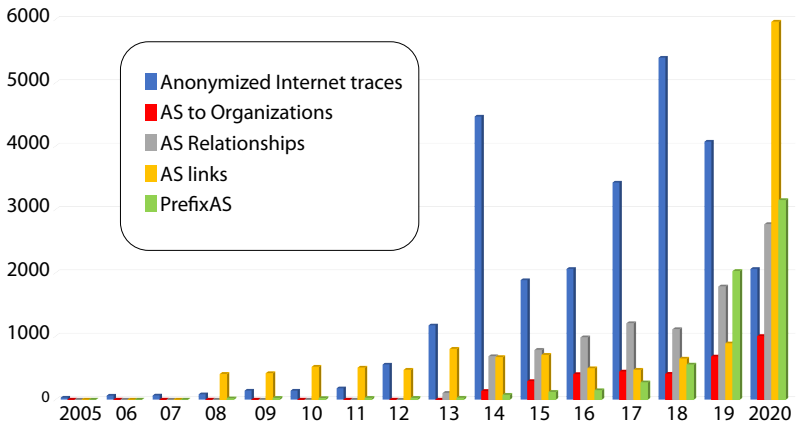

Figure 2: Data Distribution Statistics: Unique users downloading CAIDA data. Our last trace was collected in January 2019.

References

[1] C. Hall, R. Clayton, R. Anderson, and E. Ouzounis, *Inter-X: Resilience of the Internet Interconnection Ecosystem*. European Network and Information Security Agency (ENISA), April 2011. http://www.enisa.europa.eu/activities/Resilience-and-CIIP/critical-infrastructure-and-services/inter-x/interx/inter-x.

[2] 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.

[3] Jeremy Kepner and Jonathan Bernays and K Claffy and David Clark and Cary Conrad and Vijay Gadepally and Michael Jones and Robert Knake and Peter Michaleas and Chad Meiners and Robert Morris and Sandeep Pisharody and Sarah Powazek and Andrew Prout and Philip Reiner and Doug Stetson and Leah Walker, "Zero Botnets: A Defend Foward Approach (DRAFT)," 2021.

[4] J. Kepner, K. Cho, k. claffy, V. Gadepally, P. Michaleas, and L. Milechin, "Hypersparse Neural Network Analysis of Large-Scale Internet Traffic," in *IEEE High Performance Extreme Computing Conference (HPEC)*, Sep 2019.

[5] Center for Applied Internet Data Analysis, "The CAIDA Anonymized Internet Traces Dataset (April 2008 - January 2019)," 2008-2019. https://www.caida.org/data/passive/passive_dataset.xml.

[6] "Crypto-PAn: Cryptography-based Prefix-preserving Anonymization." http://www.cc.gatech.edu/computing/Networking/projects/cryptopan/.

[7] Alberto Dainotti and Alistair King, "STARDUST - Sustainable Tools for Analysis and Research on Darknet Unsolicited Traffic," 2017. https://www.caida.org/funding/stardust/.

[8] "Kentik, Inc.," 2020. https://www.kentik.com/about/.

[9] V. Giotsas, "Periscope: tool and API," 2016. http://www.caida.org/tools/utilities/looking-glass-api/.

[10] RIPE, "Using published RPKI data." https://www.ripe.net/manage-ips-and-asns/resource-management/rpki/using-published-rpki-data.

[11] Center for Applied Internet Data Analysis (CAIDA), "CAIDA Data Overview." https://www.caida.org/data/overview/.

[12] Center for Applied Internet Data Analysis (CAIDA's) Data Resource Catalog, 2020. https://catalog.caida.org.

[13] Globus, University of Chicago, "Globus." https://www.globus.org/data-sharing.

[14] Coleridge Initiative, "Rich Context Competition." https://coleridgeinitiative.org/richcontextcompetition/workshopagenda.