Project Summary

Effective Internet measurement raises daunting issues for the research community and funding agencies. Improved understanding of the structure and dynamics of Internet topology, routing, workload, performance, and vulnerabilities remain a disturbingly elusive priority, in part for lack of large-scale distributed network measurements available to scientific researchers. Ironically, even the research community networks struggle to make progress on these essential obstacles to cyberinfrastructure research. The data dearth is understandable. Measurement of operational Internet infrastructure involves navigating more complex and interconnected dimensions (logistical, financial, methodological, technical, legal, and ethical) than measurement in most scientific disciplines. CAIDA has been navigating these challenges with modest success for fifteen years, collecting, coordinating, curating, and sharing data sets for the Internet research and operational community in support of Internet science.

We propose three concrete contributions to the IRNC community's measurement efforts: to foster and distill discussion of how to best make IRNC data and statistics available, to adapt two CAIDA measurement technologies for IRNC community needs, and to experiment with two innovations in data-handling procedures applied to existing IRNC measurements. We will accomplish the first task by organizing and hosting a series of workshops, including half-day workshops at IRNC PI meetings to discuss IRNC measurement priorities and identify how CAIDA and other researchers can support them. In between IRNC PI meetings we will have 2-day annual workshops dedicated to measurement activities, to build on and extend previous efforts of the IRNC measurement group and explore in depth how the community can make better use of perfSONAR, metadata, and other data-handling and data-protection technologies.

Second, we propose to improve two CAIDA measurement technologies we already know can better serve the IRNC community: (1) We will upgrade our traffic reporting software to be IRNC-friendlier, by adding functionality to recognize IPv6 and DNSSEC, support anonymization and aggregation for privacy protection, and read data formats used by the majority of the IRNC operators, such as netflow output from routers. (2) We will (optionally) install, deploy, and manage IPv6-capable active measurement nodes at each interested IRNC site. IPv6 Internet reachability measurements are of particular interest since available data suggests the educational and government-supported communities are deploying IPv6 before the commercial sector.

Third, we propose to apply two innovations in data-handling procedures to existing IRNC measurement data. The first is a recently proposed framework for privacy-sensitive data sharing, to apply to data not appropriate for public posting, but explicitly requested through designated channels to use in clearly defined research. Second, we propose to illustrate our community building effort with a landmark reporting deliverable: a prototype of a "Bureau of Internet Statistics" report, hopefully inspiring other network infrastructure communities to join in this effort.

Intellectual merit. The proposed work will help IRNC operators better understand their networks by making more effective use of data they already collect as well as newer technologies for measurement and visibility of their nteworks. The data, tools, and distillations resulting from this effort will be made available to researchers using a privacy-sensitive sharing framework and will advance research in a number of sub-disciplines of network science.

Broader impact. Contributions from this project promise to strengthen activities in network modeling, simulation, analysis, and theoretical research, enabling the IRNC program to play a formative role in the emerging discipline of network science, and enhancing NSF's leading role in sustainable stewardship of cyberinfrastructure.

Contents

IRNC-SP: Sustainable data-handling and analysis methodologies for the IRNC networks

Project Description

1 Motivation: Improving the Lenses on IRNC Networks

Effective Internet measurement raises daunting issues for the research community and funding agencies. Improved understanding of the structure and dynamics of Internet topology, routing, workload, performance, and vulnerabilities remains a disturbingly elusive priority. The dearth is understandable. Measurement of operational Internet infrastructure involves navigating more complex and interconnected dimensions (logistical, financial, methodological, technical, legal, and ethical) than measurement in most scientific disciplines. The Internet's historical trajectory also contributes to current conditions: the world's emerging TCP/IP-speaking infrastructure escaped too early from the government-funded research and education community, into a fully deregulated environment with no concerted data collection at all.

Balancing individual privacy against other needs, such as national security, critical infrastructure protection, or even science, has long been a challenge for law enforcement, policymakers and scientists. It is good news when regulations prevent unauthorized people from examining the contents of your communications, but current privacy laws make it hard to provide academic researchers with data needed to scientifically study the Internet. To make matters worse, the few data points suggest a dire picture [1], shedding doubt on the Internet's ability to sustain its role as the world's preferred communications substrate.

Far from having scientific predictive power over the Internet's future, we must acknowledge that the Internet research community is not even in a position to thoroughly comprehend the Internet's past, since there has been no imperative (or resources dedicated) to record history. Academic network researchers today often choose scientific problems based on what data they can manage to scrape together (bottom-up) rather than choosing the most important problems to study and getting the data needed to rigorously study them [2]. New researchers typically do not even know what can be measured, much less how to navigate the policy issues in collecting and processing measurements. Even when researchers find relevant data, it is rarely in a directly usable form [3].

Upon request of a previous NSF program manager (KT) for the International Research Network Connections (IRNC) program, last year we examined the state of IRNC network monitoring statistics [4, 5, 6, 7, 8] to suggest improvements, as well as speaking with IRNC measurement group members (Matt Mathis and Matt Zekauskas) who undertook a previous evaluation [9]. Each monitoring effort presents useful graphs for individual links, but each is using different tools and presenting different data, so aggregation or comparison is difficult, as is extracting general insights into network usage and connectivity. Many measurement plans have gone unimplemented, and online statistics activities become unmaintained due to changing needs and priorities. Even with the perfSONAR framework, consistent, comparable reporting on workload, topology, routing, and performance remains lacking. This gap derives from resource constraints, knowledge gaps in what tools are available, and insufficient policy frameworks to support protected data-sharing. We try to address all three issues in the proposed work.

2 Proposed Work

We propose three concrete contributions to the IRNC community: to foster and distill discussion of how to best make IRNC data and statistics available, to adapt two CAIDA measurement technologies for IRNC community needs, and to experiment with two innovations in data-handling procedures applied to existing IRNC measurements.

2.1 Foster Community Engagement

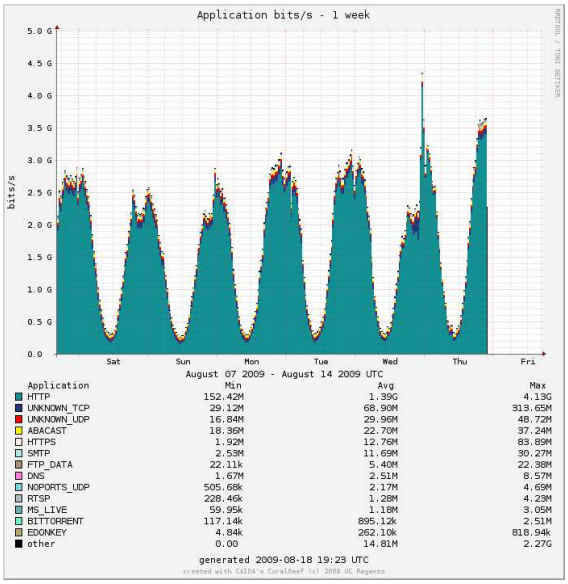
A high priority goal of this project is to expose the IRNC community to a range of provably effective measurement, visualization, and data-sharing technologies, to help them take advantage of the recent progress in this field. We propose to organize two annual workshops, one co-located with the IRNC-PI meeting if possible, and the other hosted at CAIDA, in each year of the project. The workshops will collect, synthesize, and publish feedback on data, tools, and data-sharing technologies; IRNC-PI meeting workshops will discuss IRNC measurement priorities and what CAIDA can offer to support them. We will summarize and publish reports for all workshops.

We propose half-day workshops co-located with IRNC-PI meetings during the project period. At the first half-day workshop, we will introduce the IRNC community to recent technical innovations in open source measurement tools and protected data sharing, including anonymization and metadata catalogs. Subsequent IRNC-PI meeting workshops will present the results of data gathered from newly installed tools, assist IRNC sites with interpretation of the data, solicit improvements for more effective access to the data, and discuss lessons from other projects that may help the IRNC community make better use of perfSONAR and other community frameworks for sharing data.

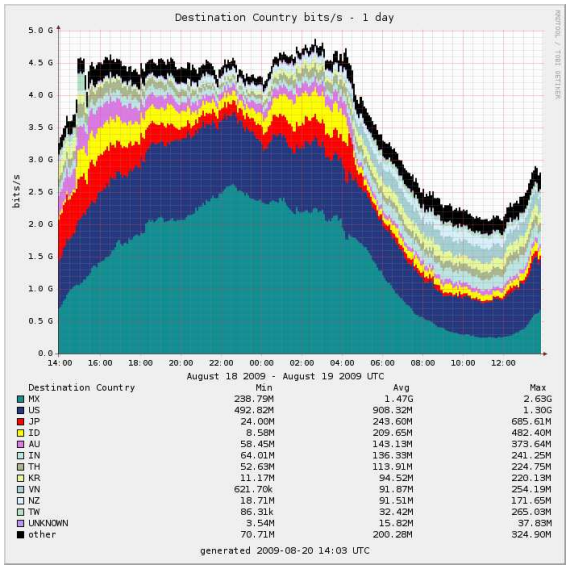
In between IRNC PI meetings we will have 2-day annual workshops dedicated to measurement activities, to build on and extend previous efforts of the IRNC measurement group [10, 11], and will specifically explore in depth and distill discussion of the following topics, for publication and presentation at the subsequent PI meeting:

1. what tools should be made perfSONAR-aware/capable and how
2. how to make finding and extracting perfSONAR data easier
3. how to make existing IRNC traffic data available for aggregation, comparison, and trend analysis
4. best practices in anonymization [12, 13]
5. augmenting existing statistics with IPv4:IPv6 usage ratio measurements over time
6. participation in existing data-sharing projects such as "A Day in the Life of the Internet" [14] by contributing meta-data for data that IRNC already collects.
7. soliciting other ideas on how to use existing measurement capabilities

Additional virtual meetings in between the PI meetings, will enable intermediate feedback on data analyses, and reaching out to other established international network research and analysis communities, including RIPE [15], EIFFEL [16], and MOMENT [17] in Europe, and WIDE [18] and APAN in Asia [19] (a sample of international groups who have requested CAIDA attendance at workshops to help advance measurement discussions). We would like to connect these research communities more effectively to the IRNC community. We also expect the IRNC community to benefit from our planned structured DatCat discussion forums to facilitate and archive discussion among data seekers and data providers.



(a) Application bits/s



(b) Destination Country bit/s

Figure 1: *Sample CoralReef traffic reporting output from an OC-192 (10GB) monitor(a) by application (currently using port numbers); (b) by country. We propose to upgrade CoralReef to handle router data formats and applications used by the majority of the IRNC operators.*

2.2 Improve IRNC Access to Innovations in Strategic Measurement Capabilities

We propose to enhance two CAIDA measurement technologies, our traffic reporting software and our topology measurement platform, to better support the IRNC community.

2.2.1 Traffic measurement: making Coral traffic reporting software IRNC-friendlier

CoralReef is a comprehensive open source software suite developed by CAIDA to collect and analyze data from passive Internet traffic monitors, in real time or from trace files [20]. Realtime monitoring support includes system network interfaces (via libpcap), and Linux and FreeBSD drivers for Endace DAG (POS and ATM) cards. The package also includes programming APIs for C and perl, and applications for capture, analysis, and web report generation. Figure 1 shows a snapshot of two examples of CoralReef reporting output. The left graph shows traffic applications over time (based on ports only)[1], and the right graph shows distribution of traffic by destination country. We propose to upgrade CoralReef to support: IPv6, DNSSEC, anonymization, aggregation, and parse input data formats used by the majority of the IRNC operators, particularly netflow output from routers, so it should work with most existing IRNC traffic measurement capability.

We recognize the state of IRNC network monitoring statistics [4, 5, 6, 7, 8] is such that immediate use of these tools is not possible in every case, but a flexible, customizable open source platform for flow measurement is a reasonable tool to add to an IRNC-accessible toolbox. At some point perfSONAR may accommodate flow data into perfSONAR, at which point we can make CoralReef perfSONAR-aware.

Figure 2: *As of mid-July 2009, there are 37 Ark monitors in 23 countries. We propose to install, deploy, and manage IPv6-capable Ark nodes at IRNC sites and distill reachability data from them.*

2.2.2 Connectivity measurement: subsidizing IRNC participation in Archipelago

CAIDA has been measuring, analyzing, modeling, and visualizing global Internet topology for over a decade. CAIDA's newest active measurement infrastructure, Archipelago (Ark), consists of several dozen standard PC's deployed around the world, running software that allows them to operate as a coordinated secure measurement platform capable of performing various types of Internet infrastructure measurements and assessments. Figure 2 depicts the 37 active Ark monitors deployed as of August 2009. We deploy approximately 10 Ark monitors per year, in geographically as well as organizationally diverse locations, to comprehensively sample the global Internet topology and conduct other global Internet measurements. We hope to expand our IPv6 measurement capabilities; currently, 8 of the deployed monitors have working IPv6 connectivity.

We propose (optionally) to install, deploy, and manage IPv6-capable Ark nodes at each IRNC site (we have provisioned budget for 5 nodes). We will distill reachability data from Ark monitors for presentation in perfSONAR-compatible format. IPv6 Internet reachability is of particular interest since available data suggests the educational and government-supported communities are deploying IPv6 before the commercial sector [21, 22] Figure 3 illustrates the high number of research networks represented in CAIDA's current AS-core map of global Internet connectivity. IRNC participation in the Archipelago cloud would improve global visibility into growing IPv6 penetration.

Some members of the IRNC community expressed regrets when NLANR's AMP service [23] was decommissioned, as it helped with network diagnostics and troubleshooting. Now an Ark node may serve as a possible provisional substitute. Ark hosting sites have found that the probing

[1]Coral uses RRD tool to provide daily, weekly, monthly, and 2-yearly views of the data, consistent with what most IRNC sites already seem to use.

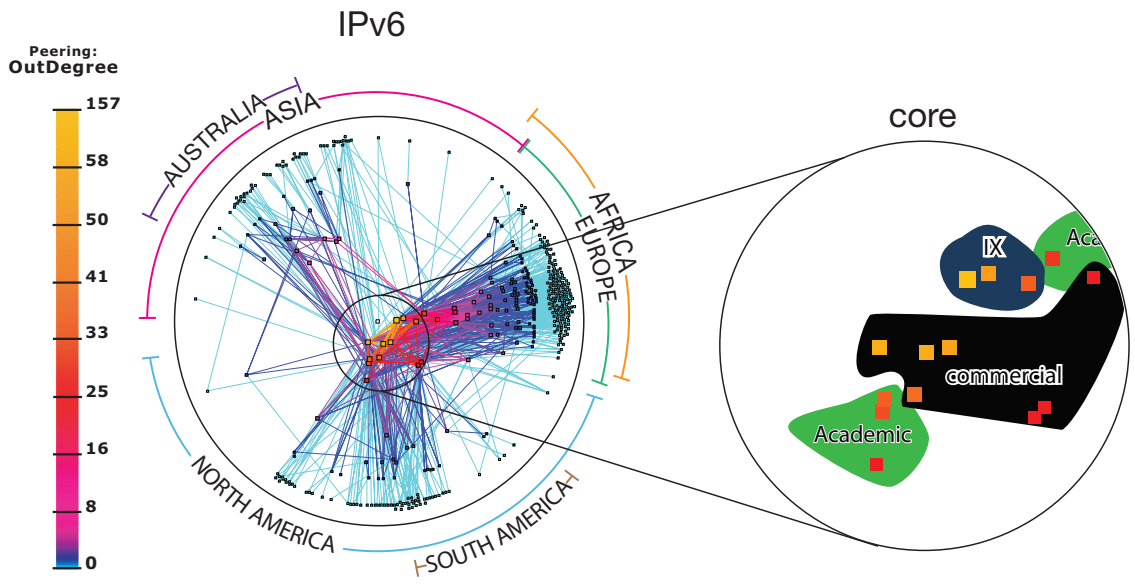


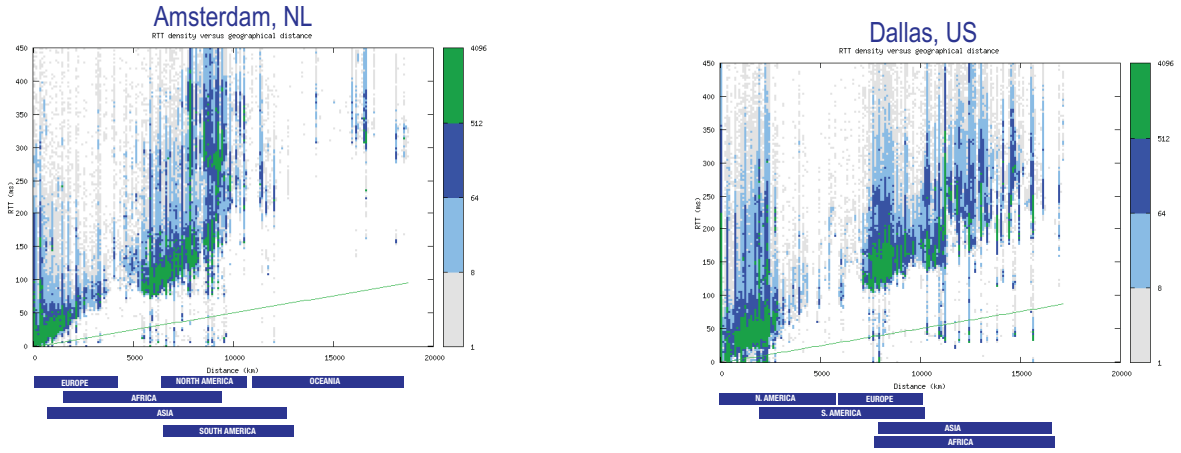
Figure 3: *In contrast to the IPv4 AS Internet "core" dominated by U.S. and multi-national commercial networks, the IPv6 AS core is a more heterogeneous mixture of European networks, exchange points, and particularly strong participation by academic institutions around the world. The latter make IRNC sites ideal locations for Ark IPv6 connectivity monitors, as they would provide insight and awareness that would benefit the IRNC community as well as the global Internet.*

application (currently scamper [24]) helps them identify routing configuration problems, similar to those mentioned in a 2006 IRNC report on international R&E routing problems [25]. To support these functions, we have developed and maintain a set of web pages showing per-node connectivity and gathered performance statistics, as exemplified in Figure 4. At the workshops we will ask for input on what other connectivity data would be valuable, and whether Ark node installation is an option of interest to IRNC sites.

2.3 Apply Innovations in Data-handling Procedures to IRNC Data

The current default, defensive posture of not sharing network data stems from the purgatory formed by the gaps in regulation and law, commercial pressures, and evolving considerations of both threat models and ethical behavior. The threat model from not sharing network data is naturally vague, as damages resulting from knowledge management deficiencies are beset with causation and correlation challenges. We lack a risk profile for our communications fabric, partly as a result of the data dearth. On one hand, fortunately, society has not received the painful blows that normally motivate legislative, judicial or policy change – explicit and immediate "body counts" or billion dollar losses.[2] On the other hand, the policies that have given rise to the Internet's tremendous growth and innovation have also rendered the entire sector opaque, unamenable to objective empirical macroscopic analysis, in ways and for reasons disconcertingly resonant with the U.S. financial sector before its 2008 melt down. The opaqueness, juxtaposed with this decade's

[2]Spam already costs the world billions of dollars a year, not counting the profit made by spammers, and legislation against spam was attempted last decade, with mixed success. Reports of losses from malware and other compromises of networked systems have been reported at over billions of dollars, but their radically distributed as well as sensitive nature prevents them from getting the front page news that inspires legislative action.



(a) Amsterdam, NL

(b) Dallas, US

Figure 4: *Round-trip time (RTT) vs. geographic distance to 637k destinations from two Ark nodes. RTT clustering reflects the geographic distances of those IP addresses from the monitor. This type of data can provide hints about possible network configuration problems or enable performance comparisons across sites.*

proliferation of Internet sustainability and stewardship issues, is a cause for concern for the integrity of the infrastructure, as well as for the stability of the information economy it supports [1].

Researchers have advocated and supported sharing data for years [26, 27, 28]. Recognizing that these purely technical efforts have had limited success in supporting needed cybersecurity research, in 2008 DHS launched the PREDICT project to formalize a policy framework for balancing risk and benefit between data providers and researchers, and is working on an update to the A policy control framework enables the technical dials to allow more privacy risk if a specific use justifies it. A classic example is how to anonymize IPAs critical to answering research questions [29] e.g., traces protected by prefix-preserving anonymization may be subject to re-identification risk or content observation risk [30], but policy controls can help data providers minimize the chances that sensitive information is misused or wrongfully disclosed [31].

Other fields offer guidance on finding a balance between privacy and science [32]. Medicine has been dealing with protection of human subjects for over a century. As a response to several disturbing experiments in the field that raised public scrutiny, in 1979 the U.S. government issued the Belmont report [33] – "Ethical Principles and Guidelines for Research Involving Human Subjects" – to establish risk-benefit criteria in the assessment of research experiments. The Belmont report also clarified the concept of *informed consent* in various research settings. (DHS hosted a workshop in May 2009 for Internet researchers to discuss creating their own "Belmont report" defining acceptable boundaries of Internet experiments and subsequent data use and sharing [31].)

We propose to apply two innovations to the IRNC community's existing data operations. First, for data that is collected but too sensitive to post publicly, we advocate applying the recently proposed privacy-sensitive Internet sharing framework [34] to select data requests from network researchers via designated channels, including CAIDA's support mailing lists, Internet2's Network Research Review Council (NRRC) [35, 36] (recently co-founded by PI Claffy and Matt Zekauskas of Internet2), and PREDICT.[3]

Second, in year 3 we propose to build on our experience and analysis in the first two years of the project to publish a landmark proof-of-concept document: a sample report from the "Bureau

[3] And other recommended by IRNC PIs

of Internet Statistics", exemplifying what we believe other critical network infrastructure communities should provide as well.

2.3.1 Privacy-sensitive Internet data-sharing framework

The proposed framework integrates privacy-enhancing technology with a policy framework that applies proven and standard privacy principles and obligations of data seekers (DSs) and data provider (DPs), and allows evaluation of data-sharing techniques along two primary criteria: (1) how they address privacy risks; and, (2) how they achieve utility objectives.

The components of our framework are rooted in the principles and practices that underlie privacy laws and policies on both the national and global levels.[4] A core principle of the framework is that privacy risks associated with shared Internet data are contagious – if the data is transferred, responsibility for containing the risk lies with both provider and seeker of data. Recognizing that privacy risk management is a collective action problem, the framework contains this risk by replicating the collection, use, disclosure and disposition controls over to the data seeker. The components include: authorization, transparency, compliance with applicable laws, purpose adherence, access limitations, use specification and limitation, collection and disclosure minimization, audit tools, redress mechanisms, oversight, quality data and analyses assurances, security, training, impact assessment, transfer to third parties, and privacy laws. External advisory groups such as the Internet2's NRRC [36] or university Institutional Review Boards (IRB's) serve a key function in the framework; at the workshops we will present lessons from our experience establishing and participating in NRRC, as well as engaging with our campus IRB, that could assist in developing a similar function for the IRNC networks.

Given the legal grey areas and ethical ambiguity around disclosure and use of network measurement data, we recommend Memoranda of Understanding (MOU)s, Memoranda of Agreement (MOAs), model contracts, and binding organizational policy as enforceable vehicles for addressing privacy risk both proactively and reactively. As part of our participation in the PREDICT project, we have have developed and published a reference set of MOUs to establish a common interpretation of acceptable sharing and reasonable practices [37]; we will help interested IRNC networks adapt these MOUs for their own needs. We will also provide sample applications to campus Institutional Review Board (IRB) [38] as described in the framework.[5]

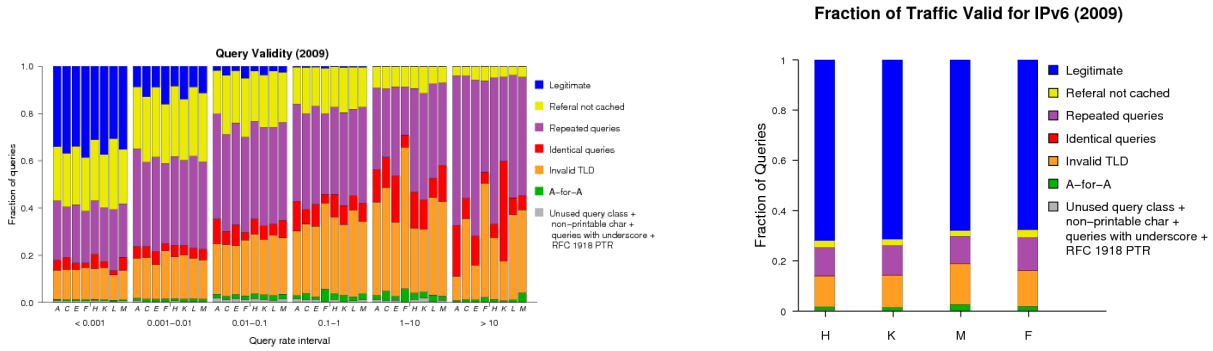
A more complete description of the framework is in [34] and will be presented at an MIT conference on "Engaging Data: First International Forum on the Application and Management of Personal Electronic Information" in October 2009. We also intend to study and report (at the second workshop) on the most challenging dimension of the data-sharing problem in an international arena: navigating the jurisdictional differences in treatment of data from various countries.

2.3.2 Prototype a "Bureau of Internet Statistics" report

We propose to establish a prototype report on the available data on network traffic, topology, routing, and performance, and security information about IRNC networks. With guidance from IRNC PIs, we will borrow from existing macroscopic infrastructure studies and reports, and ideally target dimensions of the Internet as critical infrastructure: scalability, sustainability, security and stewardship. We draw on several previous studies that include examples of the types of data

[4] In particular, the Fair Information Practices (FIPS) are considered de facto international standards for information privacy and address collection, maintenance, use, disclosure, and processing of personal information.

[5] On October 17, 2008 CAIDA submitted its first application to the UCSD campus Institutional Review Board (IRB) [38] requesting review of our traffic and other data analysis research protocol.



(a) *Query Validity* (b) *Fraction of Traffic Valid for IPv4*

Figure 5: *DNS queries to roots taxonomized by category of traffic. Although the absolute IPv6 traffic is minimal, it is notable that the level of pollution in IPv6 DNS query traffic is significantly lower.*

and analysis we envision in such a report, although we emphasize that the shape of the actual report will be a function of IRNC community consensus.

In our most recent macroscopic statistics effort, "Day in the Life of the Internet" events [14], we analyzed the largest simultaneous collection of full-payload packet traces from a core component of the global Internet infrastructure ever made available to academic researchers. This dataset consists of four large samples of global DNS traffic collected at participating DNS root servers during annual 'Day in the Life of the Internet' (DITL) experiments conducted in January 2006, January 2007, March 2008, and March 2009 [39]. We extracted historical trends, compared across data sources, and published interpretations, including growth in pollution and usage patterns that reflected ominously on the global Internet. Figure 5 shows sample DNS statistics indicating IPv6 usage patterns at the root servers for the three most recent DITL events. Although the absolute amount of IPv6 traffic is minimal, the fraction of pollution (broken or malformed queries) in IPv6 DNS query traffic is substantially lower (legitimate queries indicated by the blue bar, the rest is pollution). Similar analysis of IPv4 and IPv6 traffic (DNS or other) crossing IRNC sites would improve IRNC site awareness of relative IPv6 deployment and activity.

We include these graphs only as illustrative examples of data that can provide insight not only into one's own network, but conditions and trends on the global Internet. We propose to deliver an equivalent landmark, a prototype of an "International Bureau of Internet Statistics" report, illustrating a format that facilitates comparison and macroscopic insight into network usage, connectivity, and performance. We will borrow ideas from the Bureau of Labor Statistics [40] and its analogous agencies around the world, as well as the OECD [41].

3 Infrastructure and Collaborations Being Leveraged

We have communicated with current IRNC PI's [6], two of whom responded immediately with interest, but all of whom were slammed with a summer of proposal deadlines preventing us from meeting before the deadline. Greg Cole said he would like to dovetail this effort with his recent Taj award [42]. We also discussed our plans several times with Internet2 measurement engineers (Matt Zekauskas and Eric Boyd), including how they can enable more effective use of perfSONAR [43], which we have both offered to lead exploratory discussions at the workshops.

[6] We sent email to Greg Cole, Julio Ibarra, Jim Williams, John Sylvester, Maxine Brown, Tom DeFanti.

CAIDA has developed an Internet Measurement Data Catalog – IMDC or *DatCat* – an index of existing datasets available (under various terms) for research. DatCat tackles a significant obstacle to progress in network science: reducing the cost of searching for data by organizing information (metadata) about accessible data sets into a single repository. We have designed a new collection model to simplify the contribution and search process; we hope to leverage STCI funds (pending) to complete this and other enhancements which will be of direct utility to IRNC sites interested in sharing data. Indexing longitudinal IRNC data sets will provide three important benefits to the community: proper documentation of the data, community awareness that it exists, and an increased ability of the catalog to meaningfully respond to metadata queries. At the workshops described in section 2.1 we will offer hands-on support for IRNC operators wanting to provide metadata annotations to their data sets, including links to perfSONAR data services/lookup when appropriate.

We will be leveraging CAIDA's and UCSD's own network infrastructure as well as that of current and future hosting sites.

4 Quality of Service Metrics and Evaluation

We propose the following metrics to measure the success of each proposed contribution. To evaluate our first contribution, to *foster and distill discussion of how to best make IRNC data and statistics available*, we will solicit both structured and unstructured feedback from surveys at and between workshops, and document how this feedback will be (and was) integrated into the next year's work.

To evaluate our second contribution, to *adapt two measurement technologies for IRNC community needs*, we will keep track of the number of sites deploying Coral traffic reporting software and Ark hardware during the project period, and the number and scope of resulting data sets.

To evaluate our third contribution, to *apply two innovations in data-handling procedures*, we will uses two metrics for each innovation. For the Privacy-Sensitive Data-Sharing Framework task, we will track the number and type of sets shared with researchers under the proposed framework, and outcomes of the research.[7] For the *"Bureau of Internet Statistics" report* milestone, a summary will be submitted to a peer-reviewed journal, and available on our web site.

5 Why CAIDA is the Most Appropriate Team for this Project

Founded in 1997, CAIDA is a collaborative undertaking among organizations in the commercial, government, and research sectors aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global Internet infrastructure. CAIDA is recognized as a world leader in Internet measurement and data analysis, and has provided several landmark studies of Internet performance, workload, and topology issues [45]. CAIDA has years of experience in development, implementation, and evaluation of measurement infrastructure, as well as with

[7] Our web site has a list of known publications by non-CAIDA authors that make use of CAIDA data [44]. To summarize the 18 papers listed (a lower bound since reporting is not enforced), the AS Relationships Data has supported research on: routing on overlay networks; routing policy violations; and network security. Researchers have requested and downloaded topology data to support research in the areas of: modeling IPv4 and IPv6 AS-level topology and BGP behavior; alias resolution and router-level topology discovery; improving anycast implementations; new metrics for describing scale-free networks; evaluating router responsiveness to probes; peer-to-peer system scalability; improving visualization of complex systems; geolocation; modeling of delay; improved traceback for network attacks; and improved packet marking/filtering.

anonymization and analysis tools for the gathered data. CAIDA's long-standing trust relationships with many Internet service providers and equipment vendors facilitate monitor deployment and informed analyses. To technical, operational, research, and policy communities, CAIDA is among the most trusted sources of objective measurement tools and analyses.

CAIDA used its decade of experience with data collection, curation, and provision in developing an Internet Measurement Data Catalog to support access to network research data. We have recently proposed a Privacy Sensitive Sharing framework [34] that offers a consistent, transparent and replicable evaluation methodology for risk-benefit evaluation. In designing the framework we have considered practical challenges confronting security professionals, network analysts, systems administrators, researchers, and related legal advisors. We have also emphasized the proposition that privacy problems are exacerbated by a shortage of transparency surrounding the who, what, when, where, how and why of information sharing that carries privacy risks.

5.1 Integration of Research and Education

Some of the most valuable training for future researchers is not found in carefully controlled classroom experiences – real world data has unexpected problems. CAIDA frequently hosts graduate students from other institutions, to acquaint them with our data sets and advance collaborations. Dr. Claffy uses operational network data in lectures and seminars. CAIDA datasets are used as reference material in several courses [46, 47, 48, 49, 50] whose professors have expressed interest in additional types of data which could be supported by IRNC networks. At the workshops we will determine what packages of data should be most usefully targeted for educational use and addition to CAIDA's Internet Engineering Curriculum repository [51]. CAIDA's commitment to undergraduate education is also reflected in the 17 REU's we have supported in the last five years.

5.2 Integrating Diversity into CAIDA Activities

Based at UC, San Diego, CAIDA has a strong record of integrating diversity into our activities. Since 1999, the composition of our 90 paid interns has included 25 females, 21 Asians and 4 Hispanics. Our 25 volunteer interns in that same period have included one female, and 5 Asian students.

6 Management Plan

CAIDA personnel will be responsible for working on the proposed tasks. The requested budget supports 12.8 person-months of effort per year. The schedule of work below shows how we plan to accomplish the proposed tasks in three years of the project.

6.1 Year 1

- Conduct the 1st workshop introducing the IRNC community to available CAIDA (and other) measurement tools and techniques – 1st quarter.
- Create project web pages and start regular updates with relevant information – 1st and 2nd quarter.
- Start deploying Ark monitors at interested collaborating IRNC sites – ongoing.
- Implement proposed modifications to CoralReef software suite – 1st and 2nd quarter.
- Update report generator to include statistics of interest for the IRNC community – 2nd and 3rd quarter.

- Introduce the Privacy-Sensitive Data Sharing (PSS) Framework to IRNC members – 2nd quarter.
- Conduct the 2nd workshop to present the data gathered by newly installed tools and the analysis results – 4th quarter.

The 2nd workshop to be hosted by CAIDA at the end of Year 1 will represent a proof-of-concept demonstration of the key application concepts proposed for this project. It will summarize the progress in measurement tool deployment and application for the IRNC needs. It will set tone and define directions for the following two years of the project.

6.2 Year 2

- Publish the workshop report – 1st quarter.
- Continue deploying Ark monitors at collaborating IRNC sites – ongoing.
- Create web pages showing per-node connectivity and performance monitored by Ark at participating IRNC sites – 1st quarter.
- Organize an add-on workshop co-located with the IRNC PI meeting to report progress and discuss ongoing measurement issues – TBD (depends on the PI meetings' schedule).
- Assist the IRNC Data Providers with preparing MOUs and MOAs for their data sharing projects – ongoing.
- Continue CoralReef modifications to implement user feedback and requests – ongoing.
- Host the 4th workshop at CAIDA to assess the achieved progress and to collect ideas for the "Bureau of Internet Statistics" report – 4th quarter.

6.3 Year 3

- Publish the workshop report – 1st quarter
- Maintain Ark monitors at collaborating IRNC sites and the corresponding statistics web pages – ongoing.
- Conduct an add-on workshop co-located with the IRNC PI meeting to report progress and discuss ongoing measurement issues – TBD (depends on the PI meetings' schedule).
- Refine the Privacy-Sensitive Data Sharing Framework concept based on feedback and experience of IRNC Data Providers – 2nd quarter.
- Prepare a sample report actualizing the "Bureau of Internet Statistics" (BIS) concept – 2nd and 3rd quarter.
- Host the final workshop to discuss and finalize the BIS report – 4th quarter.

7 Results from Prior Support

1. **SCI: ITR-(NHS+EVS)-(dmc+SIM): Improving the Integrity of Domain Name System (DNS) Monitoring Trends.** SCI-0427144, $3,397,981 Sep 04 - Aug 09 (Claffy) This project addressed National and Homeland Security recommendations by the President's Critical Infrastructure Protection Board to develop a 'cyberspace network operations center (NOC)'. We accomplished the central mission of this proposal - to provide data needed to support DNS research. We conducted annual global measurement events (Day-in-the-life of the Internet – DITL) and catalogued the resulting data sets into the DatCat.
2. **NeTS-FIND Greedy Routing on Hidden Metric Spaces as a Foundation of Scalable Routing Architectures without Topology Updates** CNS-0722070, $714,998 Oct 07 - Sep 09 (Claffy and Krioukov) The proposed research involves concerted cross-fertilization across fields of

networking, theoretical computer science, physics, and mathematics. We are developing a novel network modeling methodology that is elegantly generic in nature, mathematically sound, and promises a solution to one of the most challenging problems of future large-scale networking. The GROHModel represents a rigorous mathematical foundation for truly scalable routing architectures in dynamic networks.

3. **CRI Community-Oriented Network Measurement Infrastructure.** CNS-0551542, $583,900 Sept 06 - Sep 11 (Claffy) Internet research critically depends on measurement, but effective Internet measurement raises several daunting issues for the research community and funding agencies. There is increasing awareness that obtaining a better understanding of the structure and dynamics of Internet topology, routing, workload, performance, and vulnerabilities calls for large-scale distributed network measurement infrastructure. CAIDA proposes to upgrade both of our current measurement infrastructures (passive and active) to provide the research community data from the wide area Internet that will target the need for validation of current and proposed efforts in large-scale network modeling, simulation, empirical analysis, and architecture development to answer questions of critical national security and public policy importance.

References

[1] kc claffy, "Ten Things Lawyers Should Know About the Internet." http://www.caida.org/publications/papers/2008/lawyers_top_ten/.

[2] R. Hamming, "You and your research: transcription of the Bell Communications Research Colloquium Seminar," Mar. 1986. http://www.cs.virginia.edu/~robins/YouAndYourResearch.pdf.

[3] A. Odlyzko, "Minnesota Internet Traffic Studies (MINTS)." http://www.dtc.umn.edu/mints/home.php.

[4] "TransPAC2 Statistics," 2009. http://www.transpac2.net/stats.php.

[5] "GLORIAD Monitoring System," 2009. http://www.gloriad.org/gloriad/monitor/.

[6] "Pacific Wave: Technology Measurements," 2009. http://www.pacificwave.net/technology/measurements/.

[7] "WHREN Monitoring & Measurement," 2009. http://whren.ampath.net/network/monitor.htm.

[8] "TransLight StarLight Measurement," 2009. http://www.startap.net/translight/pages/measurement.html.

[9] M. Mathis and M. Zekauskas, "IRNC International Measurement Group – A Survey of Current Projects Along With Recommendations for Improvement," 2006. https://wiki.internet2.edu/confluence/display/IMP/2006+Interim+Report.

[10] Internet2, "Internet2 International Measurement project - IRNC Measurement Group," 2008. https://wiki.internet2.edu/confluence/display/IMP/IRNC+Measurement+Group.

[11] Internet2, "Internet2 international measurement project wiki," 2008. https://wiki.internet2.edu/confluence/display/IMP/Home.

[12] CAIDA, "Summary of Anonymization Best Practice Techniques." http://www.caida.org/projects/predict/anonymization/.

[13] CAIDA, "Bibliography of Papers on Internet Data Anonymization." http://www.caida.org/publications/bib/networking/bytopic/index.xml#anonymization.

[14] CAIDA, "A Day In The Life of the Internet," 2009. http://www.caida.org/projects/ditl/.

[15] RIPE, "Réseaux IP Européans Network Coordination Centre (RIPE NCC)," 2009. http://www.ripe.net.

[16] "The EIFFEL Initiative, a Support Action (SA) proposed for the 7th Framework Programme (FP7).." http://www.eiffel-thinktank.eu/.

[17] "Monitoring and Measurement in the Next Generation Technologies (MOMENT)." http://www.fp7-moment.eu/.

[18] J. Murai, "Widely Integrated Distributed Environment." http://www.wide.ad.jp/.

[19] "Asia Pacific Advanced Network." http://apan.net/.

[20] D. Moore and K. Keys, "CoralReef Software Package." http://www.caida.org/tools/measurement/coralreef/.

[21] kc claffy and Bradley Huffaker and Young Hyun, "ARIN and CAIDA IPv6 Survey Summary – April 2008," 2008. http://www.caida.org/publications/presentations/2008/arin_survey/.

[22] kc claffy and CAIDA, "ARIN and CAIDA IPv6 Survey Summary – October 2008 ," 2008. http://www.caida.org/publications/presentations/2008/arin_survey_summary/.

[23] "Active Measurement Project." http://amp.nlanr.net/.

[24] M. Luckie and CAIDA, "scamper," 2008. http://www.caida.org/tools/measurement/scamper/.
[25] C. Robb and J. Williams, "International R/E Routing," 2009. http://www.transpac2.net/documents/2006/International.Routing.v1.0.doc.
[26] M. Allman, V. Paxson, and T. Henderson, "Sharing is caring: so where are your data?," *ACM SIGCOMM Computer Communication Review*, vol. 38, Jan 2008.
[27] Vern Paxson, "The Internet Traffic Archive." http://ita.ee.lbl.gov/.
[28] CAIDA, "Internet Measurement Data Catalog Project." http://www.datcat.org/.
[29] S. Coulls, C. Wright, A. Keromytis, F. Monrose, and M. Reiter, "Taming the Devil: Techniques of Evaluating Anonymized Network Data," in *Proceedings of the 15th Annual Network and Distributed Systems Security Symposium*, Feb. 2008.
[30] Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," Tech. Rep. Legal Studies Research Paper No. 09-12, University of Colorado Law, Aug. 2009. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.
[31] K. Claffy, "What's Belmont Got To Do With It?," june 2009. http://blog.caida.org/best_available_data/2009/06/12/whats-belmont-got-to-do-with-it/.
[32] K. Claffy, "The inevitable conflict between data privacy and science," Jan. 2009. http://blog.caida.org/best_available_data/2009/01/04/the-inevitable-conflict-between-data-privacy-and-data-utility-revisited/.
[33] N. I. of Health, "Ethical principles and guidelines for the protection of human subjects of research (the belmont report)," Apr 1979. http://ohsr.od.nih.gov/guidelines/belmont.html.
[34] E. Kenneally and K. Claffy, "An Internet Data Sharing Framework For Balancing Privacy and Utility," in *Engaging Data: First International Forum on the Application and Management of Personal Electronic Information (to be presented)*, Oct. 2009. http://www.caida.org/publications/papers/2009/engaging_data/.
[35] K. Claffy, "Internet2 Launching its own IRB," Jan. 2009. http://blog.caida.org/best_available_data/2008/10/10/internet2-launching-its-own-irb/.
[36] Internet2, "Network Research Review Committee," 2009. http://www.internet2.edu/networkresearch/nrrc.html.
[37] CAIDA and PREDICT, "PREDICT Memoranda Of Agreement (MOA)," 2007. http://www.caida.org/projects/predict/mou/.
[38] CAIDA, "Caida's application to the ucsd irb," 2008. http://www.caida.org/home/about/irb/caida_irb_app_cover_17oct2008.xml.
[39] S. Castro, D. Wessels, M. Fomenkov, and k claffy, "A Day at the Root of the Internet," *ACM SIGCOMM Computer Communications Review*, Oct. 2008.
[40] U. G. Department of Labor, "Bureau of Labor Statistics," 1884. http://www.bls.gov/.
[41] OECD, "Internet Infrastructure Indicators," 1998. http://www.oecd.org/dataoecd/11/25/2091083.pdf.
[42] G. Cole and J. Sobieski, "The Taj: A New Model for Global Federated Network Infrastructure for Science and Education," Aug. 2009. http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0943314.
[43] 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.
[44] "Papers Published Using CAIDA Datasets." http://www.caida.org/data/publications/.

[45] CAIDA. http://www.caida.org/outreach/papers/.
[46] J. Rexford, "Computer Networks Course CS461," 2006. http://www.cs.princeton.edu/courses/archive/spr06/cos461/.
[47] A. Bhattacharya, "Computer Networks Course CS584," 2009. http://www.cs.nmsu.edu/~amiya/cs584/.
[48] M. Caesar, "Advanced Internetworking CS598," 2008. http://www.cs.uiuc.edu/homes/caesar/classes/CS598.F08/.
[49] Dr. Cedric, "Wednesdays with Dr. Cedric: Network Performance, Modeling and Monitoring," 2009. http://cs259upd.blogspot.com/.
[50] S. Faber, "Network Situational Awareness (Course at CMU)," 2009. http://www.andrew.cmu.edu/course/95-855/.
[51] CAIDA, "Internet Engineering Curriculum Project." http://www.caida.org/projects/iec/.