DIBBS EI: Platform for Applied Network Data Analysis (PANDA)

For the last 20 years CAIDA has developed many data-focused services, products, tools and resources to advance the study of the Internet, which has permeated disciplines ranging from theoretical computer science to political science, from physics to techlaw, and from network architecture to public policy. As the Internet and our dependence on it have grown, the structure and dynamics of the network, and how it relates to the political economy in which it is embedded, is gathering increasing attention by researchers, operators and policy makers, all of whom bring questions that they lack the capability to answer themselves. CAIDA has spent years cultivating relationships across disciplines (networking, security, economics, law, policy) with those interested in CAIDA data, but the impact thus far has been limited to a handful of researchers. The current mode of collaboration simply does not scale to the exploding interest in scientific study of the Internet.

In response to feedback from these communities, we propose to create a new platform offering researchers more accessible, calibrated and user-friendly tools for collecting, analyzing, querying, and interpreting measurements of the Internet ecosystem. First, we will integrate existing research infrastructure measurement and analysis components developed by CAIDA that, once connected, will enable new scientific directions, experiments and data products for a wide set of researchers. We will emphasize efficient indexing and processing of terabyte archives, advanced visualization tools to show geographic and economic aspects of Internet structure, and careful interpretation of displayed results. Our second task will be active engagement of collaborators from four targeted disciplines: networking, security, economics, and public policy. To achieve this goal, we will: organize annual workshops to establish and stimulate multi-disciplinary collaborations; develop online video tutorials targeting non-networking experts as well as classroom-focused materials; maintain an annotated bibliography and discussion forum; and institute an advisory board to provide strategic direction. Our third task focuses on platform extensibility and adaptability to new opportunities. By Year 3 we will integrate into the platform new data products that target unmet research needs: (1) a comprehensive DNS data set that will facilitate mapping network behavior to a human view; (2) anonymized residential network traffic measurement data, a wholly distinct type of data that will inspire an emerging generation of Internet data scientists.

Collaborating Partners. Our 27 collaborators are well-known experts in the targeted disciplines. Colleagues at UCSD will contribute expertise in science gateway development, big data analytics leveraging HPC resources, and familiarity with the DIBBS community. Aben (RIPE-NCC), Bailey (UIUC), Beverly (NPS), Caesar (UIUC), Crovella (BU), Clark,Lehr (MIT), Deccio,Zappala (BYU), Deri (NTOP/Pisa), Dimitropoulos (ETH), Donnet (Liege), Frieden (PSU), Giovannetti (U.Anglia), Greenstein (Harvard), Haverkort (Twente), Johnston (FCC), Jordan (UCI), Katz-Bassett (USC), Luckie (Waikato), Smith (Penn), Gartzke, Gupta, Roberts, Smarr, Snoeren, Voelker, Wilkins-Diehr (UCSD).

The **intellectual merit** of this project lies in the new empirical studies possible in the four targeted disciplines. The proposed platform will address NSF's CIF21 goal of interconnecting cyberinfrastructure components to develop a comprehensive, robust, scalable shared resource that will bridge diverse communities, and integrate HPC, data, software, and facilities to expand the potential of Internet-related science. The platform will trigger innovations in: Internet mapping and path prediction; detection of route hijacking and other disruptive events; cybersecurity preparedness; economic studies of correlations between ISP characteristics, market power, performance degradations, security practices, and regional economic growth; and regulatory discourse that has thus far occurred largely without data.

Broader impacts of this project will include increased public awareness about Internet structure, dynamics, performance, and evolution, and will inform discussion of critical issues in current and future large-scale networking. The developed tools will lower the threshold to use CAIDA's data products and platforms for R&E needs.

Contents

DIBBs: EI: Integrated Platform for Applied Network Data Analysis (PANDA)

1 Introduction: Expanding the Reach and Impact of Internet Infrastructure Data

For two decades, UC San Diego's Center for Applied Internet Data Analysis (CAIDA) has been developing data-focused products, services, open source software tools, and resources to advance the field of Internet science, which has permeated disciplines ranging from theoretical computer science to political science, from physics to tech law, and from network architecture to public policy. As the Internet and our dependence on it have grown, the structure and dynamics of the network, and how it relates to the political economy in which it is embedded, is gathering increasing attention by researchers from all those disciplines, as well as network operators and policy makers, all of whom bring questions that they lack the capability to answer themselves. Epistemological challenges lie in developing and deploying measurement instrumentation and protocols, expertise required to soundly interpret and use complex data, lack of tools to synthesize different sources of data to reveal insights, and privacy issues. As it has become clear how many fears and aspirations about the Internet (security, affordability, neutrality, universal service, congestion) are rooted in economics, ownership, and trust issues, CAIDA has cultivated communities of economists, law, and policy researchers with an interest in empirically grounding their understanding of the Internet. But the impact thus far has been limited to a handful of researchers. The current mode of collaboration simply does not scale to the exploding interest in scientific study of the Internet, nor to complex and visionary scientific uses of CAIDA's data by non-networking experts.

In response to feedback from these communities, as well as our own insights from meetings, workshops, and other discussion forums, we propose to create a new shared cyberinfrastructure resource – the *Platform for Applied Network Data Analysis (PANDA)*. To create this system we will use existing research infrastructure measurement and analysis components, that, once connected, will enable new scientific directions, experiments, and data products. The idea for this platform builds on input from dozens of researchers over 7 years of workshops, and our design plan supports its use, evaluation, extensibility, and sustainability. The proposed system will integrate active Internet measurement capabilities, multi-terabyte data archives, live data streams, heavily curated data sets revealing coverage and business relationships, and traffic measurements that represent, for many researchers, the holy grail of scientific sources of information about Internet behavior. Our goal is to enable a broad set of researchers to access, query, visualize, and analyze Internet data, as well as create new data products in ways that promote valid interpretations of data and derived inferences. We will also develop new visualization tools to allow non-experts to understand various aspects of Internet structure, using geographic and economic annotations on the data, with access controls where appropriate for sensitive data.

Our second task will be active and responsive engagement with four fields via our 27 collaborators, with the goal of expanding the use of data for multi-disciplinary research, including in Internet mapping and path prediction; detection of attacks and other disruptive events; cybersecurity preparedness; economic studies of correlations between ISP characteristics, market power, performance degradations, security practices, and regional economic growth; and empirical grounding for emerging regulation. We have budgeted staff time to support these collaborators, workshops to conduct existing and stimulate new collaborations; development of online video tutorials, an annotated bibliography and moderated discussion forum; and an advisory board to provide strategic direction and community stewardship.

To demonstrate extensibility and adaptability of the platform to new opportunities, our third task is to integrate into the platform two new data building blocks that target unmet research needs per the attached letters of collaboration: (1) home network traffic measurement; and (2) a comprehensive mapping of the DNS namespace. Quantitative evaluation metrics will include usage statistics, survey responses, time to integrate new modules, and downloads of workshop materials, reports, and online resources.

This Early Implementation project targets DIBB's primary goals: *enabling new data-focused services, capabilities, and resources to advance scientific discoveries, collaborations, and innovations.* Per the solicitation's expectations, PANDA will *build upon, integrate with, and contribute to existing community measurement infrastructure and annotated data collections*, and offers *synergy with current national policy priorities in cybersecurity, including privacy, information technology research ethics, and interoperability of relevant data building blocks.* Further, this project will bring together cyberinfrastructure expertise and researchers in multiple disciplines, to ensure that the platform addresses researcher needs, especially cross-disciplinary data interpretation challenges.

Section 2 (§2) describes the existing building blocks that we propose to integrate into the PANDA platform, including their features and limitations. Our first task (§3) is to create and maintain PANDA – aimed at ultimately inspiring an innovative, reliable, sustainable and accessible ecosystem of software and services to advance scientific inquiry about the Internet [1]. We have split this task into three p arts: development work to scale performance functional requirements of existing components; development work to create hooks between platform components, and integration of components behind a flexible *science gateway* web interface. Our second task (§4) is a community-driven approach customized to the four targeted disciplines, to ensure constructive use, evaluation, and evolution of this software platform. Task 3 (§5) extends this evaluation to collaborator-driven testing of the ability to incorporate new components. We selected components for their potential to address unanswered scientific research questions in four disciplines (§6), and to promote synergies among these disciplines.

2 Existing Data Building Blocks to Integrate

Figure 4 (the included system diagram) presents an overview of the initial components of PANDA, which all represent recognized data collections and tools we share with the community. CAIDA supports and shares many other measurement tools and supporting libraries [2], as well as data collection platforms [3] we hope to integrate into the PANDA system in the future. We selected components for the first version that are most synergistic with the research agendas of our collaborators.

2.1 Archipelago Active Internet Measurement Platform and Supporting Components

Archipelago, Vela, Henya. Archipelago (Ark) [4] is CAIDA's (CRI-funded) active measurement infrastructure [5] running software that allows distributed nodes to operate as a coordinated secure measurement platform. Since 2007, Ark has gathered the largest global Internet topology data for use by academic researchers, from (as of January 2017) 170 Ark monitors located on six continents and hosted by diverse organizations: research/educational, commercial, network infrastructure, residential, etc.[1] Vetted researchers can run experiments directly on the nodes [7, 8, 9, 10, 11, 12, 13, 14, 15], via an Ark-provided API, or using our **interactive web interface Vela [16]**. We also continuously gather our most comprehensive and scientifically generative active data set [17] by systematically measuring IP-level paths to a dynamically generated list of IP addresses covering the entire routed IPv4 and IPv6 address space. Since 2007 we have gathered 48B traceroutes (over 20 TB) as well as DNS lookups of all IP addresses observed during probing. With growth in monitor deployment and use (and growth of the Internet topology), we project these datasets to grow by 19B traces (approximately 10 TB) per year.

Few researchers can download data sets of this size, so to facilitate discovery of the full potential value of this raw data, we have created an **interactive web-based exploratory interface – *Henya*** [18] to allow researchers to find the most relevant data for their research, such as all traceroutes through a given region and time period toward or across a particular address, autonomous system (AS), or country. (One network may operate one or more autonomous systems, depending on engineering and business practices, e.g., mergers. There is no official data base mapping AS numbers to organizations owning and operating them; CAIDA maintains a heuristic-based mapping [19].) We continue to develop analysis and visualization modules for Henya to support study of specific events as well as longitudinal trends in performance and structural evolution. Henya's user interface is targeted at networking experts, so we created an educational tutorial video [20] explaining how to use Henya in terms non-experts can understand.

Curated targeted data sets. Some research questions do not require the raw data, so (also based on community feedback) we regularly publish heavily curated data sets that researchers use widely [21, 22, 23, 24]. We curate two-week snapshots of raw traceroute data into Internet Topology Data Kits (ITDK) [25] (Figure 1) which provides inferred, annotated router-level and AS-level topologies of the global Internet. We

[1] Map and details of Ark monitor coverage: http://www.caida.org/projects/ark. In 2012, we ported our measurement software platform to the Raspberry Pi [6].

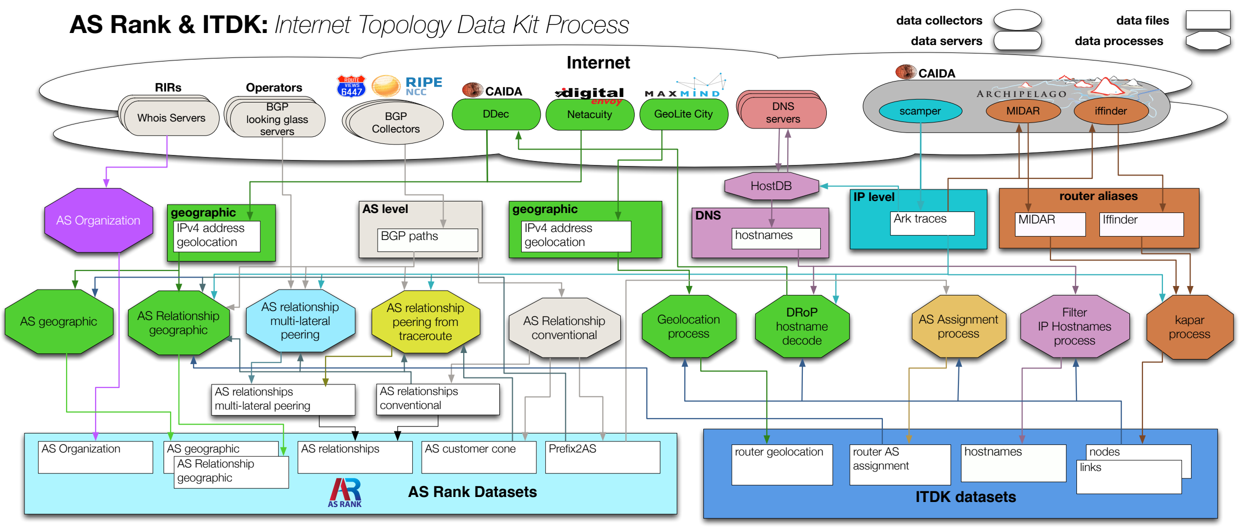


Figure 1: Internet topology data measurement, mining, and analysis lifecycle process for data CAIDA collects and/or curates. PANDA will strive for a more consistent and user-friendly accessible interface to these data sets.

have increased their richness over time by integrating new techniques as we develop them, including AS ownership inference [26] and scalable alias resolution (identifying which interface IP addresses belong to the same routers), which is required to convert the IP-level topology discovered by traceroute to a router-level topology [27]. To inform selection of topology datasets for specific research needs, we have published analyses comparing different sources of topology data [28, 29, 17] for constructing AS- and router-level graphs [30, 31, 32]. Based on feedback from a survey of ITDK users, researchers agree on its potential power but find its utility limited (especially for non-experts) by its daunting complexity with no user interface.

2.2 AS Rank: Comparison of routing and economic relationships among ISPs

AS Rank (as-rank.caida.org) [33] is CAIDA's interactive web interface that allows one to explore routing and business relationships between ISPs (identified as ASes in the routing system) and organizations that own them. The Internet AS-level topology and its dynamics are consequences of business decisions that Internet players make, and accurate knowledge of AS business relationships is relevant to both technical and economic forces driving Internet structure and evolution. AS relationships introduce a non-trivial set of constraints on the paths over which Internet traffic can flow, with implications for network robustness, traffic engineering, measurement strategies, and economic modeling of topology. Our AS ranking algorithm, which builds on decades of work in this area [34, 24], ranks ASes by their *customer cone* size, which is the number of their direct and indirect customer networks, as inferred from publicly available routing data. Our workshop participants over the years have consistently suggested that a more accessible and more functional user interface would greatly amplify the utility of this data.

2.3 BGPStream: efficient framework for routing (BGP) data analysis

BGPStream is CAIDA's recently developed open-source software framework for the analysis of both historical and real-time Border Gateway Protocol (BGP) measurement data. Although BGP is a crucial operational component of the Internet infrastructure, and is the subject of research in the areas of Internet performance, security, topology, protocols, economics, etc., until BGPStream was developed, there was no efficient way for researchers to process large amounts of distributed and/or live BGP measurement data. BGPStream fills this gap, enabling efficient investigation of events, rapid prototyping, and building complex tools and large-scale monitoring applications (e.g., detection of connectivity disruptions or BGP hijacking attacks). We have demonstrated how to apply components of the framework to different scenarios, as well as how to deploy complex services with BGPStream [35, 36]. Integration of BGPStream into our other data platforms would improve their performance, scalability, and functionality.

2.4 Periscope: Extending measurement coverage by leveraging operational infrastructure

Although Ark has 170 well-distributed nodes, it sees only a fraction of Internet connectivity. Researchers have expanded coverage by leveraging *looking glass* (LG) monitoring infrastructure that network operators deploy themselves to support remote execution of non-privileged diagnostic tools, such as traceroute, ping or BGP commands, through web interfaces [37, 38, 39, 40, 41, 42]. However, lack of input and output standardization has hindered development of automated systematic use of LGs. We developed *Periscope*, a middleware API that automates discovery of and interaction with looking glass server nodes around the world (2,591 LG nodes across 562 ASes, as of December 2016) [43, 44]). Periscope combines crowd-sourced and cloud-hosted querying mechanisms for scalability, including Virtual Machines (VMs) in the Google Compute Cloud (GCC) and Amazon Elastic Compute Cloud (EC2) platforms to execute HTTP requests to the underlay Looking Glasses (LGs). An intelligent controller coordinates execution of queries so as to honor the various LG querying rate limitations. Preventing abuse is important, not only ethically but also because overwhelming the LGs would likely lead to their decommissioning from public use. We initially developed this tool to support our research in mapping peering interconnections to a facility [45], but we maintain the system for the benefit of the research community, and share the source code on request. However, supporting users is still a rather manual process, mostly to prevent users from gaming the system by registering multiple user accounts. Also, Periscope is not connected to any other measurement or data products, limiting its potential.

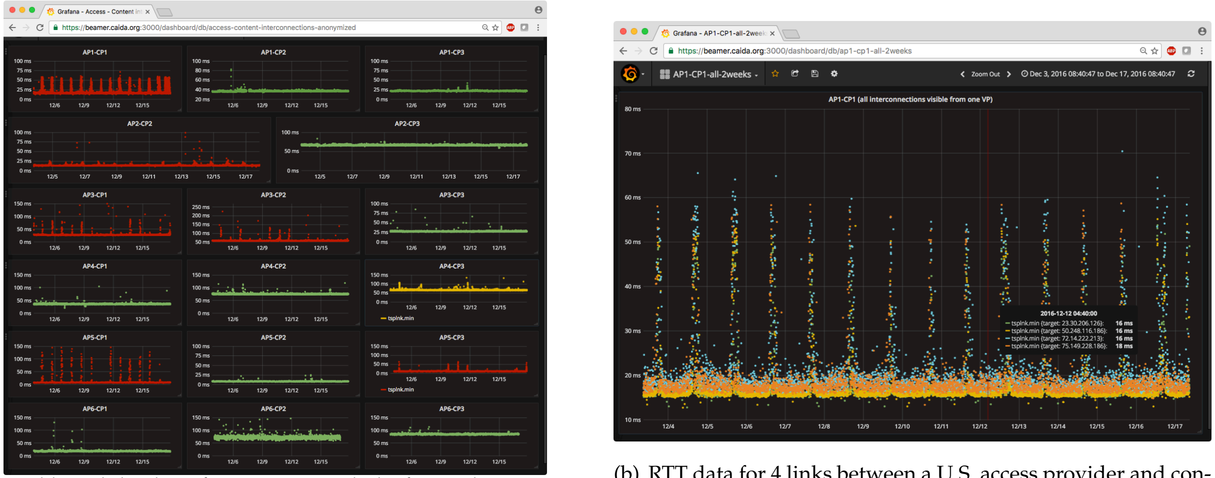
2.5 MANIC: Mapping and Analysis of Interdomain Congestion

With shifting market power in the Internet ecosystem, links connecting access providers to their peers, transit providers and major content providers have become a potential point of discriminatory treatment. While the FCC has regulatory authority over those links, they have acknowledged they lack sufficient expertise to develop appropriate regulations thus far.[2] As part of a NeTS project (§7), we developed a system to measure and analyze congestion on interconnection links [48]. The core idea of our technique is to send (TTL-limited) probes from a vantage point (VP) within a network, toward the near and the far end of an interdomain (border) link of that network, and to monitor diurnal patterns in the near and far-side time series. A persistently elevated RTT to the far end of the link, but no corresponding RTT elevation to the near side, is a signal of congestion at the interdomain link. The more daunting challenge is to identify these interdomain border links via active measurements from the edge of a network, which we also tackled for this project, with our current method implemented in our open-source *bdrmap* tool [49].

We developed a backend system that manages this probing from (in January 2017) 60 Ark (and 15 Bismark [50]) VPs, collects and organizes data, and presents that data for analysis and visualization. To adapt to routing changes, we run *bdrmap* continuously on each VP, and update each monitor with knowledge of which IP addresses a VP should probe to observe interconnection links of interest. The system pulls the probing data from VPs and indexes it into an influxDB time-series database with a Grafana front end[3] within 30 minutes of being generated. Figure 2(a) shows a pre-configured dashboard that plots the recent congestion state of interdomain links from 5 large access networks in the U.S. to 3 interconnecting content providers/CDNs. Plots with a pronounced diurnal pattern signal evidence of congestion on that interdomain link. Figure 2(b) shows a week of RTT data for 7 interdomain links between a U.S. access provider and a content provider. This graph took a few minutes to produce; previously this data exploration would have taken hours using data stored in standard relational databases.

[2] In response to concerns about such discriminatory treatment, the U.S. Federal Communications Commission (FCC) attached conditions to the 2015 AT&T-DirectTV merger that require AT&T report to the FCC performance measurements of its interconnections [46], in part to develop the FCC's understanding of how to measure such discrimination in the first place. CAIDA served as the Independent Measurement Expert for these conditions [47].

[3] InfluxDB can support millions of time series, scalable and usable read/write API, and SQL-like querying capability. Grafana integrates seamlessly with influxDB to provide interactive querying and graphing capability. The plugin can process a stream of data as it is pushed into influxDB and perform user-defined processing to generate alerts.


Grafana - Access - Content In
https://beamer.caida.org:3000/dashboard/db/access-content-interconnections-anonymized
AP1-CP1
AP1-CP2
AP1-CP3
AP2-CP2
AP2-CP3
AP3-CP1
AP3-CP2
AP3-CP3
AP4-CP1
AP4-CP2
AP4-CP3
AP5-CP1
AP5-CP2
AP5-CP3
AP6-CP1
AP6-CP2
AP6-CP3
tsplink.min
Grafana - AP1-CP1-all-2weeks
https://beamer.caida.org:3000/dashboard/db/ap1-cp1-all-2weeks
AP1-CP1-all-2weeks
Zoom Out
Dec 3, 2016 08:40:47 to Dec 17, 2016 08:40:47
AP1-CP1 (all interconnections visible from one VP)
80 ms
70 ms
60 ms
50 ms
40 ms
30 ms
20 ms
10 ms
12/4
12/5
12/6
12/7
12/8
12/9
12/10
12/11
12/12
12/13
12/14
12/15
12/16
12/17
2016-12-12 04:40:00
tsplnk.min {target: 23.30.206.126}: 16 ms
tsplnk.min {target: 50.248.116.186}: 16 ms
tsplnk.min {target: 72.14.222.213}: 16 ms
tsplnk.min {target: 75.149.228.186}: 18 ms

(a) Dashboard display of congestion on links from 6 large access networks in the U.S. to 3 content providers

(b) RTT data for 4 links between a U.S. access provider and content provider.

Figure 2: PANDA will enable linking of MANIC data with current topology data to allow querying of specific links in the topology, or visibility of topology surrounding link displayed in a specific graph.

2.6 Spoofer: Assessment of IP address validation best practices (LOC: Waikato)

The set of security-relevant data one might integrate into PANDA is vast, and we can imagine a subsequent phase of this project focused on such data. As a proof-on-concept, we chose a security measurement and reporting platform with urgent policy relevance and implications: CAIDA's DHS-funded Spoofer project [51]. In collaboration with U. Waikato (see LOC), CAIDA has developed new open-source client-server tools that enable crowd-sourced measurements of source address validation (SAV) compliance [52, 53]. Source address validation is a quintessentially incentive-incompatible configuration best practice that protects other networks from spoofed denial-of-service attacks coming from one's own network. 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. We have developed client software that works on Windows, MacOS, and UNIX-like systems, periodically testing a network's ability to both send and receive packets with forged source IP addresses (spoofed packets). We have started to produce reports and visualizations of results, for the benefit of operators testing their own configurations, management trying to justify IT resource allocation, and remediation authorities who want to prioritize SAV compliance attention where it will most benefit. Combining this data and other security-relevant data (blacklists) and industry structure data would allow systematic correlation of properties that seem to hinder or promote cybersecurity preparedness.

3 Task 1: PANDA: Platform for Applied Network Data Analysis

With popular as well as scientific attention to Internet measurement growing, and based on feedback from collaborators overwhelmed with the complexity and range of CAIDA's resources, our goal with this project is to offer a smooth transition between several existing CAIDA interactive interfaces and data products, which leverage a well-tested data sharing framework for researchers using the data [54]. Based on many discussions with different target communities, we have a clear vision of what we need to do to make the building blocks described in §2 more useful to them. First, many components need re-architecting to **scale performance** to a larger number of users, which will require software development and systems integration to leverage existing HPC resources at SDSC (§3.1). Second, we need to develop APIs and software modules to create **hooks between platform components**, which we will design and implement in a coordinated way to support two modes of use: (1) integration of components behind a flexible **science gateway** web interface that accommodates different levels of expertise; (2) development of one's own scripts and tools, so users with more advanced programming skills can take advantage of the upgraded components (§3.2).

Third, we have a long list of feedback on how to make the presentation of different data more accessible to non-experts, which we will use to frame development of the science gateway (§3.3).

3.1 Software development to scale performance and functionality for community use

Some of our data platforms we built as research prototypes, and they are not capable of keeping up with the increasing resource demands being placed on them as they grow in popularity for research as well as classroom use. We are already using XSEDE (Gordon and Comet) resources for BGPStream and other projects at CAIDA [55], so we will be able to build on this experience.[4]

1. **Re-architect AS Rank to use HPC resources on the back end to improve scalability and performance**. AS Rank has trouble meeting current demands, especially when professors decide to use it for interactive classroom use.[5] We will re-architect AS Rank from the ground up to use XSEDE resources on the back end to support parallelization of real-time calculations on large topology graphs (hundreds of thousands of links). This task will include developing new intelligent indexing schemes such as those used in Henya (for IP-level path information) to create a highly efficient AS path database backend. A more efficient, HPC-based design will allow computations not possible with the current system, e.g., show all paths visible for a given AS, show which observed paths informed a given relationship inference. It will also allow tracking changes to AS paths, relationships, and customer cone sizes over time. We will extend the database to compute and store not just the size of a customer cone, but the set of ASes in the cone, to enable comparisons of customer cone diversity across providers and regions. Finally, we will improve AS-level visualizations to highlight structure from the perspective of a given AS, which users find engaging.
2. **Unify scalable border mapping techniques (LOC Penn).** In collaboration with UPenn, we will combine mapping techniques from our own *bdrmap* [49] and UPenn's MAP-IT [56] tools (both recently published in the same session at IMC). This development will allow us to create a unified border mapping module which researchers can interactively run on large archives of traceroute data, including RIPE Atlas and Periscope, to increase observation of locally scoped peering.
3. **Scale up the MANIC functionality to integrate additional vantage points, border mapping, and geolocation capabilities.** The MANIC platform currently uses Ark vantage points to measure congestion on interdomain links of a network *from within that network*. MANIC's archive of performance data is currently at 8TB of data, increasing at about 100GB/month, but interest in expanding this platform will amplify this growth. We will integrate Periscope and RIPE Atlas vantage points to provide a view of an interconnection link from the other direction, or in some cases to allow comparative views of a given interconnect from different locations. Even more exciting, the U.S. FCC has also asked us (but has no funding to support) to expand our data collection infrastructure to include the FCC-SamKnows (Measure Broadband America) deployment of home routers, which provides thousands of additional VPs in the U.S. from which to measure interconnection. This expansion would increase storage requirements to 1TB/month. Additionally we will integrate geolocation information to enable systematic analysis of congestion inferences between networks [57], including integration of our facility-level information [45], i.e., in which building an interconnection occurs. We will rearchitect the influxDB database instance to work on a cluster of XSEDE nodes, so that it can scale to multiple cores and support many concurrent users and queries. Fortunately, the tens of thousands of time-series we accumulate by measuring performance across each observed interconnect are highly amenable to parallelization.

3.2 Create software modules to link components to each other and external software

Researchers from each targeted community have executed or requested three classes of use of our data products: (1) selection of specific snapshots of raw data to download for further analysis (as in Henya); (2)

[4] We have used Gordon specifically to support real-time reporting of UCSD telescope traffic, which requires aggregating traffic data in different ways (e.g., count of unique source IPs per country) and extract millions of time series metrics that are stored in time series databases on the Gordon IO node, and updated 24/7. If the telescope continues operation, we will pursue separate funding sources to integrate this more sensitive data into PANDA in the future.

[5] ASRank's current aging database server has 128GB RAM, 6-core 3.2GHz CPU, and 1.7TB non-SSD RAID.

use of an interactive dashboards for simple queries and visualization of topology (AS Rank), performance (MANIC), or security data (Spoofer); or (3) modules for researchers to write analysis programs that will interface with PANDA components, like the InfluxDB suite of tools to support user-defined functions for stream or batch data processing from the database backend (which MANIC uses). This third mode could allow CAIDA staff to efficiently turn research questions posed by non-technical colleagues into programs that are realistic to run in practice. For each mode of use, we have already architected at least one existing component to efficiently support that mode of use, so we can make an informed estimate of the work required for other modules. As with Task 1, we will be able to leverage expertise, and sometimes bits of software, across different projects at CAIDA.

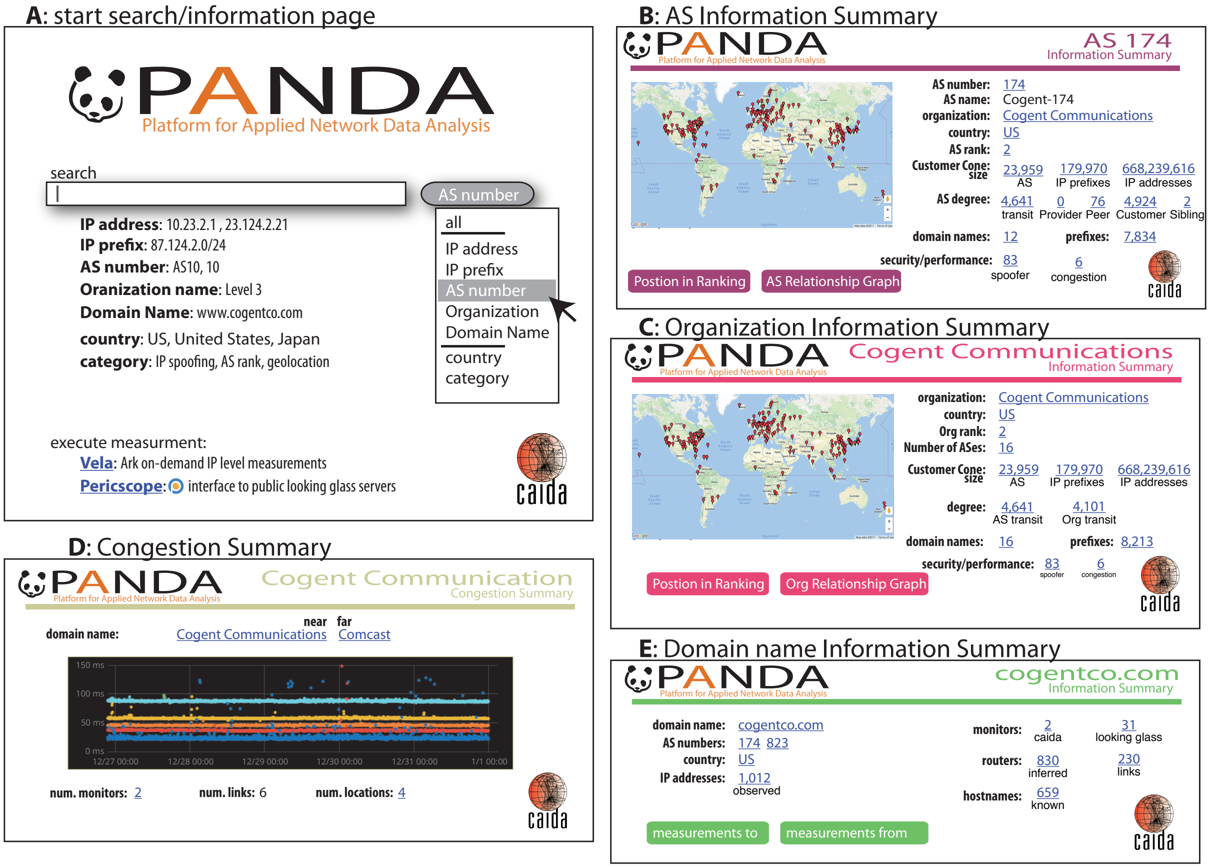
1. **Implement links between archived and real-time measurements, and between AS-level and IP-level measurements.** As Archipelago evolves[6] we will enable users to easily transition from querying archived measurements to requesting on-demand measurements in real-time, or vice-versa. That is, users will be able to compare historical to current characteristics of paths. We will also add support to link an IP-level view to an ISP (AS)-level view, so that when one executes a measurement, it will be easy to see all archived and derivative data related to all networks crossed across the probed path.
2. **Integrate the use of BGPStream consistently for all of PANDA's components.** We will implement bindings to the main BGPStream C library to facilitate use by external software modules, e.g., for route hijacking and outage detection. We will make available distribution-specific packages to simplify installation, including Ubuntu, Debian, FreeBSD, and CentOS.
3. **Enable Periscope to leverage other software platforms.** For faster resource allocation, tighter control of network resources, and resilience to failures, we will enable Periscope to use both Ark and RIPE Atlas nodes as part of its querying underlay. Intelligent allocation of querying can conserve the Periscope querying budget for use with measurements that *only* its underlying LGs can satisfy, We will also integrate Periscope as a BGPStream broker [35] to enable consumption of Periscope BGP data through the BGPStream APIs.[7]
4. **Cross-correlate MANIC data output to other measurement projects and data sets.** Policy makers who interpret congestion data also care about the business relationships between ASes. We will integrate AS-relationship data from our AS-rank system to allow users to slice the congestion data according to the inferred business relationship between networks. Other projects are collecting different sorts of data that can provide evidence of performance degradations [57], e.g. the video quality reports and throughput/latency measurements from the FCC Measuring Broadband America platform [58]. We will explore how to compare such data sets to enrich our view of interconnection, while also making the data more accessible and usable to researchers and policy makers.

3.3 Increase community accessibility of unified platform and underlying components

The recent launch of the Science Gateways Community Institute (SGCI) [59] at SDSC was a Ëureka" moment for us, as we realized that science gateways are an ideal model for how we can expand the scientific and collaborative opportunities of our community data resources across more research domains. A *science gateway* [60] is an application-specific web-based resource for accessing HPC resources for data, software, and services that let users focus on the science, rather than the complexities of accessing unfamiliar data formats and HPC systems. Science gateways can also enable well-formed communities, such as CAIDA and its collaborators, to focus application development on tools and access methods rather than the on the logistics of procuring, curating, and managing data and supporting systems. CAIDA's facilities are housed in the San Diego Supercomputer Center (SDSC), which is pioneering this type of cyberinfrastructure, and we are fortunate to have SGCI PI Nancy Wilkins-Diehr (LOC) and her team collaborating on this project to support access to more backend (XSEDE) resources, data restructuring to optimize performance, and usability review. CAIDA presented the AS Rank and Henya interfaces at the Gateways 2016 event [61], where we also conceived of our proposed design of interactive query and visualization functionality for PANDA. Figure 3 shows a mockup front-end and sample interactive screens of the proposed science gateway, which

[6] The CRI project (§7) covers expanding Ark's footprint, support for new measurements, and browsing, querying, and visualization functionality specific to traceroute data.

[7] Although Periscope BGP measurements are publicly available, consumption of historical measurements currently requires use of the API, which makes it hard to combine views with other BGP sources.



A: start search/information page

search

IP address: 10.23.2.1 , 23.124.2.21
IP prefix: 87.124.2.0/24
AS number: AS10, 10
Oranization name: Level 3
Domain Name: www.cogentco.com
country: US, United States, Japan
category: IP spoofing, AS rank, geolocation

execute measurment:
Vela: Ark on-demand IP level measurements
Pericscope: interface to public looking glass servers

B: AS Information Summary

AS 174 Information Summary

AS number: 174
AS name: Cogent-174
organization: Cogent Communications
country: US
AS rank: 2
Customer Cone size: 23,959 AS 179,970 IP prefixes 668,239,616 IP addresses
AS degree: 4,641 transit 0 Provider 76 Peer 4,924 Customer 2 Sibling
domain names: 12 **prefixes:** 7,834
security/performance: 83 spoofer 6 congestion

C: Organization Information Summary

Cogent Communications Information Summary

organization: Cogent Communications
country: US
Org rank: 2
Number of ASes: 16
Customer Cone size: 23,959 AS 179,970 IP prefixes 668,239,616 IP addresses
degree: 4,641 AS transit 4,101 Org transit
domain names: 16 **prefixes:** 8,213
security/performance: 83 spoofer 6 congestion

D: Congestion Summary

Cogent Communication Congestion Summary

domain name: **near** Cogent Communications **far** Comcast

num. monitors: 2 **num. links:** 6 **num. locations:** 4

E: Domain name Information Summary

cogentco.com Information Summary

domain name: cogentco.com
AS numbers: 174 823
country: US
IP addresses: 1,012 observed
monitors: 2 caida 31 looking glass
routers: 830 inferred 230 links
hostnames: 659 known

Figure 3: Mockup of Panda interface. Users would start with screen **A** *, enter a query, and progress to a context-based screen based on query type, e.g., if they specify AS 174, they will see an interface similar to* **B**. *Underlined words represent hyperlinks to pages with additional details. For example* **B**'*s congestion 6 link (lower right of B) would take the user to screen* **D** *which would show information for congestion on six links observed by PANDA components.*

allows the user to choose a specific ISP, company, region, IP address, or domain name of interest. and then select from a set of possible data products related to that object. The initial screen on the upper left (**A**), will take users to different screens (samples **B** through **E**), depending on their request.

Although we already have experience in re-architecting individual components, and with linking pairs of components (e.g., Spoofer and AS Rank), thus far we have not done it with usability by non-technical experts in mind. Our guiding motivation for PANDA is to facilitate exploratory data analysis and interoperability across passive and active measurement data sources, as a means of connecting a broad range of scientific users to data that could transform their research, and to other researchers that can help them interpret it. PANDA is a first step toward support for much richer query and analysis elements that could support an even broader range of sophisticated research questions in the future. Our initial goals are:

1. **Improve usability of the ITDK.** Based on feedback from several of the target communities, we will create a simplified version that removes complex artifacts in the data, e.g., multiple origin ASes, AS loops and sets, hyperlinks [31], in order to render it amenable to processing by basic graph database tools. We will provide documentation to explain inferential implications of the simplifications. We will also create an ëconomist-friendly" version annotating the data in different ways (company names in addition to network numbers) to support economics and policy use of the data.
2. On request, **provide data products in additional easier-to-use, domain-specific formats**; for example, JSON, standard graph formats, and input formats of network simulators. We may either offer our data converted to these forms or provide custom tools for doing so and thus reduce the impedance mismatch that currently hinders data exchange and research.
3. **Create user-friendly interface to source address validation (SAV) compliance information accessible to operators and policy-makers via the PANDA gateway.** First, we will integrate published

information on which networks are present at each IXP to report the SAV compliance state of networks at the IXPs where we have test results. This information will allow IXP members to adopt defensive practices against non-compliant networks if it chooses to be present at the IXP. Second, we will integrate public BGP information on the stability of edge network address space, to enable analysis of the feasibility of deploying static access control lists as a form of compliance with SAV best practices. [53].[8]

4 Task 2: Support for and collaboration with multiple disciplines

The most externally visible cyberinfrastructure output of this project will be the integrated PANDA system and its capabilities. But, as the HCI saying goes, *"A design isn't finished until somebody is using it."* A design that works for a technical expert may not work for an economist or policymaker. One of our creative innovations is to leverage the significant cyberinfrastructure investments needed to support networking and security research, so that it can also support SBE with long-standing important questions and little means to answer them. This task has five activities: staff time dedicated to supporting collaboration; annual workshops that focus on user engagement with the PANDA system and associated data sets; annual community surveys of all PANDA users to solicit feedback on usability and impact of the platform; development of online video tutorials suitable for classroom use, and a moderated forum for discussion of related empirical studies; and an advisory board to provide strategic direction and community stewardship.

1. **Dedicated staff time to support collaborations with economists and policymakers.** We have allotted significant staff and PI time to enrich existing and generate new collaborations among policy analysts, economists and measurement experts, and using these collaborations to evolve the design of user interface components of PANDA to better serve those targeted communities.
2. **Workshops focused on engagement with PANDA system.** Our two annual workshop series – AIMS in its 8th year and WIE its 7th – have established solid communities of collaborators in technical and social science fields, respectively. The Active Internet Measurement Systems workshop series (funded through Feb 2018 by CRI) is a forum for stakeholders in Internet active measurement projects, typically *networking and security researchers*, to explore technical and policy challenges and opportunities to maximize the scientific and operational benefit of deployed infrastructure and gathered measurements [62, 63, 64, 65, 66, 67, 68, 69]. The Workshop on Internet Economics (WIE) series brings together researchers, *Internet service providers (ISPs), economists, regulators, lawyers, and other stakeholders* to inform and debate current and emerging policy issues [70, 71, 72, 73, 74, 75]. We will expand these workshop series, and promote more cross-fertilization and collaboration across the communities. We will use these workshops to stimulate use of PANDA in multi-disciplinary collaborations, introduce new capabilities of PANDA, including hands-on tutorials to engage users, and share experiences with classroom use of PANDA.
3. We will distribute **a written survey**, at the workshops and via a larger panda-interest@caida.org community email list, soliciting feedback on presented capabilities and plans, and publish anonymized summaries of the surveys each year. We will also socialize the new platform at other community workshops and conferences where the lack of empirical data is a recognized gap.
4. We will develop and maintain an **online community resource of material**, including tutorials on how to use PANDA and its components, summaries of data-gathering efforts from different stakeholders, and a wiki/forum for independent analyses of (sometimes contradicting) results of these studies, their limitations and implications for research on measurement, economics, policy, and future network architectures. We will specifically target a portion of this resource for integration into course curriculum and student projects.
5. An **external advisory board** led by David Clark at MIT will enrich linkages between PANDA and targeted communities, identify emerging national and international issues that merit empirical attention, and suggest data sets and analysis to inform policy-making. This board will pursue vehicles such as NSF's RCN program to support outreach to other communities, e.g., TPRC, and will develop paths to sustainability beyond the DIBBS project, including possible foundation support.

[8] Static ingress access control lists work well "when the configuration is not too dynamic" [53].

5 Task 3: Extensibility and adaptation to new opportunities

To maximize chances of success, the initial components we integrate into the PANDA system will be CAIDA-developed. We will demonstrate extensibility and adaptability to new opportunities by attempting to integrate new data infrastructure building blocks from external researchers. We propose two use cases based on enthusiastic interest from and discussions with collaborators.

Comprehensive DNS measurements (LOC U.Twente). Roland van Rijswijk-Deij's team plans to integrate a comprehensive dataset that consists of large scale active measurements of the global Domain Name System (DNS) [76]. In the OpenINTEL project, jointly operated with SURFnet and SIDN Labs, U. Twente currently measures over 60% of the DNS name space once every 24 hours. By adding such comprehensive DNS data to PANDA, it becomes possible to associate network behavior observed at the IP protocol level with the human view on the Internet in the form of domain names, which helps all targeted communities accelerate inquiry as well as pursue new questions (since DNS data can change over time and researchers have no other means to map historical topology data to historical DNS data from the same time).

Home traffic measurement (LOC U.Pisa). Thus far PANDA components have focused on the structure and dynamics of the Internet, rather than user traffic measurement. Collection of user traffic data has technical challenges, but even more daunting privacy challenges. We have had extensive discussions with Luca Deri, developer of network traffic monitoring software ntop [77], about collaboratively developing and integrating a traffic monitoring tool to support monitoring in home networks, by non-technical users (using CAIDA's privacy-sensitive sharing framework [54]). Integrating BGP-aware and IXP-aware functionality (data sets and correlation capabilities) with a traffic monitoring suite will enable answering immediately relevant security questions, like has an IoT device in my home become part of a botnet?, as well as policy questions regarding evolution of the global Internet, e.g., how much of my traffic goes further than my ISP and its direct peering partners? (§6.4). Because this is a completely different type of data, we expect that integrating this component will require work on the platform itself.

6 Benefits to Research Communities

The project will achieve NSF's CIF21 goal of interconnecting cyberinfrastructure components to develop a comprehensive, robust, scalable shared resource that will bridge diverse communities, and integrate HPC, data, software, and facilities to expand the potential of experimental, Internet-related science and engineering. The proposed platform will enable the following scientific research and education opportunities. (See LOCs for more details.)

6.1 Networking Research Community

The networking research community has relied on CAIDA's measurement and data infrastructure for decades, as evidenced by our statistics on data usage (reported annually e.g., [78]) and papers published using our data [79] (including our own [80]). Our networking collaborators for this project forsee that the proposed interconnection of components will greatly facilitate their current research and enable entirely new investigations. First, the challenge of mapping the interdomain Internet has proven far greater than researchers anticipated, and the most daunting obstacle continues to be robust, scalable instrumentation to achieve both comprehensive sweeps of Internet address space and coordinated precision probing to algorithmically infer interconnection details. Many proposed infrastructure development tasks will support higher fidelity and real-time mapping capabilities, improved graph coverage, analysis, and validation, and curation of associated data sets. For example, recent work in detecting Internet exchange points (IXP) in traceroute paths [81] (FORTH) provides a powerful new tool for users, operators and researchers trying to study or troubleshoot network infrastructure. This work requires synthesizing IP, BGP, and IXP data sources into a single measurement module, but will be a trivial module to write and integrate into the PANDA system.

Other opportunities include: detection and richer characterization of MPLS tunnels and middlebox behavior, which has implications for Internet performance and user privacy (U. Liege); identification of root causes of large-scale events in the routing system (BU); reactive measurement experiments using USC's PEERING testbed (USC); disaster preparedness, assessment and recovery; and path prediction (USC,ETH) [82, 83, 84]. Integration of DNS data with BGP and traceroute data has already facilitated improved methods for geolocation of Internet infrastructure [85, 86, 87, 88, 89, 90], but enabling a more iterative research process will accelerate advances in this challenging domain. We also plan to use PANDA to uplevel our own analysis of grey market transfers (purchasing and selling) of IPv4 address blocks [91], by detecting anomalous changes in topology, DNS, and BGP data to infer address transfers.

6.2 Security and Stability Research Community

Security researchers often have to manually merge disparate data sets (or write tools to do so), if they are lucky enough to acquire them. CAIDA originally developed the BGPStream platform to provide real-time evidence of disruptive events, including attacks, on the interdomain routing system. Combining BGPStream data with AS-level meta-data in the PANDA platform will reveal insight into the nature of the attack, e.g., what types of networks are involved, or other paths through the affected networks. The ability to trigger a measurement (e.g., traceroute) or computation (e.g., latency-inferred distance, customer cone) based on input from another stream of data (e.g., interdomain routing hiccup) will enhance the capabilities to detect and mitigate route hijacking [92] and other cyberattacks (LOC ETH). The DNS data source of PANDA (LOC Twente) will facilitate mapping suspicious activity at IP-level to the human view represented by domain names. This will help our political scientist collaborators, who intend to use PANDA's capabilities to study censorship (LOC UCSD-Roberts, UIUC-Caesar), and model the impact of offensive cyberwarfare behavior (e.g., hostile disruption of connectivity) between nation-states (LOC UCSD-Gartzke).

Several of our collaborators are eager to use PANDA functionality to improve their empirical assessments of network security hygiene (Waikato), SSL certificate notary deployment, and cybersecurity preparedness (BYU-Deccio). These topics are also interdisciplinary with research on public policies related to security. For example, by correlating source address validation test results (ISPs that do not comply with BCP38/84) with network region, type, and size, policy analysts can assess how to best focus remediation policies, and understand the impact of imposed (technical or policy) measures. If this use of PANDA is successful, it will motivate extensions to study other vulnerability mitigation measures. Because security-related data is often especially sensitive, one collaborator (BYU-Zappala) is enthusiastic about extending his own SATC project to develop browser-based methods to support a content-based security model for access to sensitive components of PANDA's data repository. This collaboration could pave the way to an entirely new forms of operational security cooperation.

6.3 Economics Research Community

The industrial organization of the Internet ecosystem evolves rapidly, and PANDA will provide accessible empirical data to track changes in: patterns of interconnection (e.g., peering and direct interconnection); regional variation in industry size and modes of interconnection (public vs. private); trends in firm growth (using routing table coverage as a proxy); correlations between network business type [93] and cybersecurity preparedness (are larger companies more likely to follow security best practices? LOCs BYU); and correlations between firm size and interconnection performance (does market power influence prevalence of congestion? LOC MIT). With their own data on economic growth (which may be amenable to future integration into PANDA), economics researchers will study the relation of network infrastructure development and economic growth (LOC Harvard).

One trend received pointed interest at WIE2016: evidence suggests that the next phase in the evolution of the Internet is the integration of the traditional consumer-facing public Internet with other TCP/IP-based services that are distinct from the public Internet. Integrating our proposed collaborative development of an enhanced traffic monitoring platform (LOC Deri NTOP) into PANDA will open a new research landscape of challenging and relevant questions, such as whether most Internet traffic is shifting onto private interconnects with large CDNs directly connected to access providers [94] (LOC MIT, PSU, Anglia). This trend may imply that our traditional understanding of a globally interconnected transport platform is evolving toward regional fragmentation, with providers replicating service across regions, to the point where only the "heavy tail" of (largely non-commercial, under-capitalized) content relies on global Internet reachability, with significant implications for industry structure and regulation. To explore this hypothesis, one collaborator (LOC MIT) might use Periscope and Ark to probe paths toward major content providers over time, and evaluate trends in the extent to which replicated sites are only reachable regionally.

6.4 Public policy and legal research community

This year (2017) is critical time with respect to telecommunications regulation. After a decade of deregulatory administrative and judicial action, in 2016 the FCC acknowledged that the infrastructure and its usage had evolved sufficiently to justify reclassification of broadband back into a telecommunications service, although emphasizing its intent to forbear from most associated regulatory apparatus, at its own discretion. With the new administration, there are calls to reverse this decision, and (in the extreme) abolish the FCC or severely limit its authority [95]. Given the uncertainty of the current state of U.S. Internet policy, we find this targeted field most in need of accessible measurement and explanations of data [96, 97]. Many policy thinkers, from academic to government to industry, now consider the development of new legislation of Internet transport services inevitable, and if there is any chance of informing it with empirical data, now is the time for that effort [98, 99]. Our policy research collaborators (LOC FCC,UCI) including economists with policy interests (LOC MIT,Harvard,PSU) will help articulate policy implications of network and measurement results in formats accessible to economists and lawyers.

For example, around the world, regulatory priorities are shifting from consumer broadband access to ISP interconnection, about which there is little reliable data, leading to a proliferation of data and reports that require careful interpretation (and in some cases dismissal) of conflicting claims by stakeholders [97]. Even identifying (much less meaningfully measuring) interconnections is difficult [49], and is further complicated by merger-induced *sibling networks*, i.e., owned by the same organization, since many more paths cross sibling links and we do not reason cleanly about them. PANDA will initially support simple queries, e.g., "show me all observed interconnection links with the highest average persistent congestion in the U.S. over the last month." But working with our collaborators (LOCs FCC,UCI,MIT), we will design an interface that presents measurements structured within the context of our evolving conceptual model of interconnection [97], embedded in a deep understanding of both the technical and policy boundaries. For example, we will label interdomain links with geolocation, in some cases interconnection facility, and business relationships. PANDA will thus not only help policymakers examine this kind of data, but help them understand *how* to examine it to avoid misinterpretations, and hopefully enable the community to develop better measurements to mitigate the contradictions inherent in current data sets.

6.5 Collaborations and Synergies among Communities

Synergies among Disciplines. This project leverages CAIDA's two long-cultivated workshop-fostered communities (networking/security and economics/policy), to expand existing and stimulate new national and international collaborations to address grand challenges of cyberspace.

Other cyberinfrastructure data-sharing projects. CAIDA helped to establish and sustain the DHS IMPACT project, which offers a community of data providers and researchers legal support and vetting for access to privacy-respective sharing of sensitive Internet measurement data. IMPACT is another community of potential cyberinfrastructure experts and domain collaborators.

Synergies Across Campus and Cyberinfrastructure community. This project will generate collaborations with other projects at SDSC (LOC Gupta), the Science Gateway Community (LOC Wilkins-Diehr), and an existing DIBBS project (Pacific Research Platform at UCSD (LOC Smarr).

Education Synergies. Twenty of our 27 collaborators are teaching faculty. Task 2 includes dedicated workshop sessions to share experiences with PANDA use in curriculum and student projects.

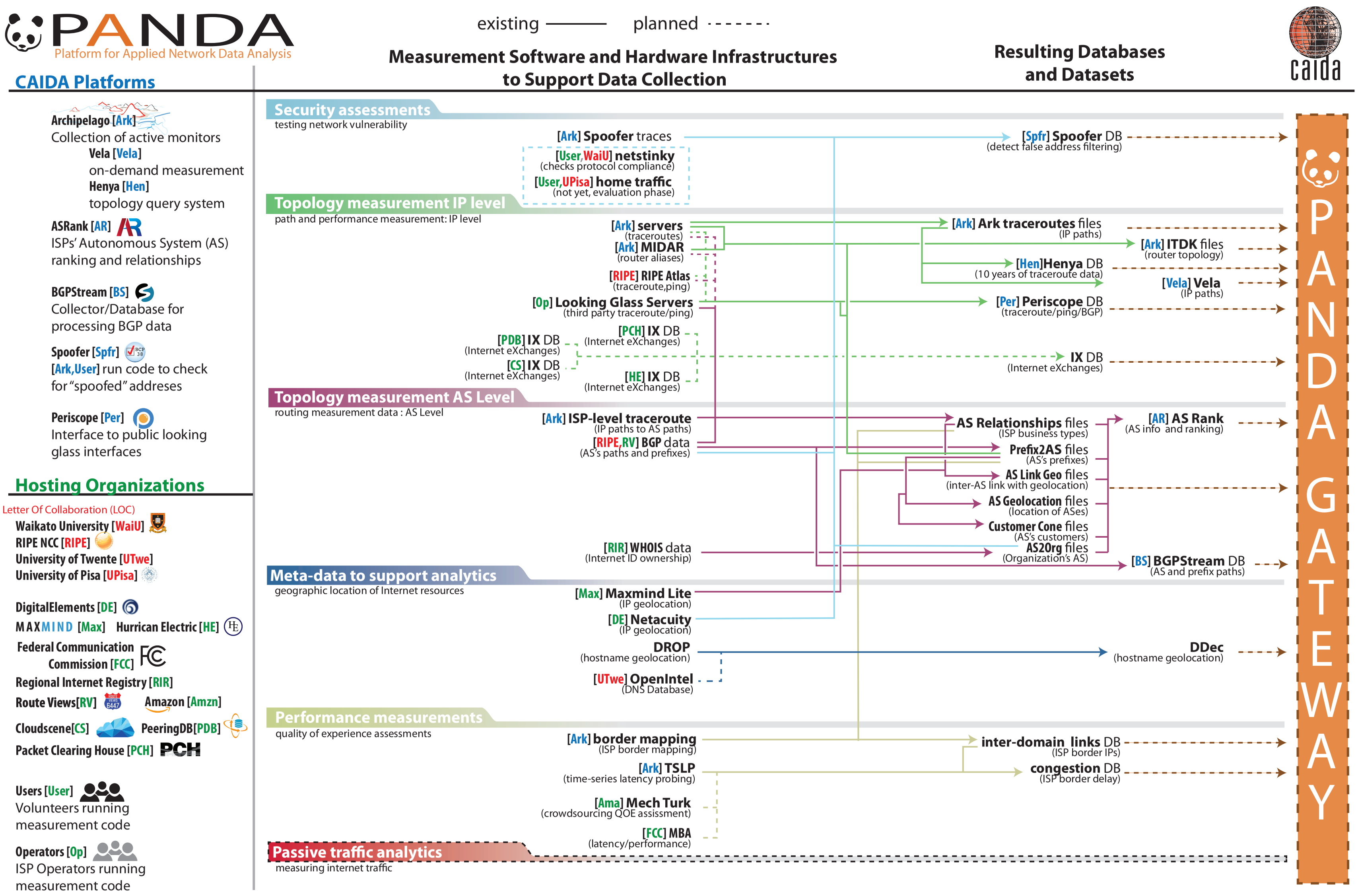


Figure 4: PANDA project components: data infrastructure building blocks

References

[1] National Science Foundation, “A Vision and Strategy for software for science, engineering, and education cyberinfrastructure framework for the 21st century,” 2012. NSF 12-113.

[2] “Caida tools - overview of caida software tools.” http://www.caida.org/tools/.

[3] Overview of CAIDA’s Data. http://www.caida.org/data/.

[4] Center for Applied Internet Data Analysis, “Archipelago Measurement Infrastructure.” http://www.caida.org/projects/ark.

[5] Center for Applied Internet Data Analysis, “Macroscopic Topology Measurements.” Research Project. http://www.caida.org/projects/macroscopic/.

[6] http://www.raspberrypi.org/.

[7] * M. Luckie, Y. Hyun, and B. Huffaker, “Traceroute probe method and forward IP path inference,” in *ACM SIGCOMM Internet measurement Conference (IMC)*, Oct 2008.

[8] * P. Mérindol, B. Donnet, J.-J. Pansiot, M. Luckie, and Y. Hyun, “MERLIN: MEasure the Router Level of the INternet,” in *Euro-nf Conference on Next Generation Internet (NGI)*, June 2011.

[9] * K. Keys, Y. Hyun, M. Luckie, and k. claffy, “Internet-Scale IPv4 Alias Resolution with MIDAR,” *IEEE/ACM Transactions on Networking*, vol. 21, Apr 2013.

[10] * R. Beverly, W. Brinkmeyer, M. Luckie, and J. Rohrer, “IPv6 Alias Resolution via Induced Fragmentation,” in *Passive and Active Network Measurement Conference (PAM)*, Mar 2013.

[11] * M. Luckie, R. Beverly, W. Brinkmeyer, and k. claffy, “Speedtrap: Internet-scale ipv6 alias resolution,” in *ACM SIGCOMM Internet measurement Conference (IMC)*, Oct 2013.

[12] * P. Marchetta, W. de Donato, and A. Pescapé, “Detecting third-party addresses in traceroute traces with IP timestamp option,” in *PAM*, pp. 21–30, Apr. 2013.

[13] * M. Luckie and k. claffy, “A Second Look at Detecting Third-Party Addresses in Traceroute Traces with the IP Timestamp Option,” in *Passive and Active Network Measurement Workshop (PAM)*, vol. 8362, pp. 46–55, Mar 2014.

[14] * R. Beverly, A. Berger, Y. Hyun, and k. claffy, “Understanding the efficacy of deployed Internet source address validation filtering,” in *ACM SIGCOMM Internet measurement conference (IMC)*, 2009.

[15] * kc claffy, “CAIDA participation in IPv6 day,” June 2011. http://blog.caida.org/best_available_data/2011/06/05/caida-participation-in-ipv6-day/.

[16] “Vela: On-Demand Topology Measurement Service.” http://www.caida.org/projects/ark/vela/.

[17] Center for Applied Internet Data Analysis (CAIDA), “The IPv4 Routed /24 Topology Dataset.” http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

[18] Young Hyun, “Henya: large-scale Internet topology query system,” 2016. http://www.caida.org/tools/utilities/henya/.

[19] B. Huffaker, K. Keys, M. Fomenkov, and K. Claffy, “AS-to-Organization Dataset.” http://www.caida.org/research/topology/as2org.

[20] Y. Hyun, “Henya: CAIDA’s Internet Topology Query System Tool,” 2016. https://www.youtube.com/watch?v=jg7CgLCMtgY.

[21] Center for Applied Internet Data Analysis (CAIDA), “AS links.” http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml.

[22] Center for Applied Internet Data Analysis (CAIDA), “Prefix to AS mappings.” http://www.caida.org/data/routing/routeviews-prefix2as.xml.

[23] Center for Applied Internet Data Analysis (CAIDA), “AS Taxonomy.” http://www.caida.org/data/active/as_taxonomy/.

[24] CAIDA, “AS links annotated with AS relationships dataset.” http://www.caida.org/data/active/as-relationships/index.xml.

[25] CAIDA’s Macroscopic Internet Topology Data Kit (ITDK). http://www.caida.org/data/active/internet-topology-data-kit/.

[26] * B. Huffaker, A. Dhamdhere, M. Fomenkov, and k. claffy, “Toward topology dualism: Improving the accuracy of AS annotations for routers,” in *Passive and Active Network Measurement Conference (PAM)*, Apr. 2010.

[27] * K. Keys, “Internet-Scale IP Alias Resolution Techniques,” *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 40, pp. 50–55, Jan 2010.

[28] H. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and A. Venkataramani, “iPlane: an information plane for distributed services,” in *Symposium on Operating Systems Design and Implementation (OSDI)*, 2006.

[29] R. Oliveira, “UCLA’s IRL Internet Topology Collection,” July 2009. `http://irl.cs.ucla.edu/topology/`.

[30] * P. Mahadevan, D. Krioukov, M. Fomenkov, B. Huffaker, X. Dimitropoulos, and kc claffy, “Lessons from three views of the internet topology: Technical report,” tech. rep., UC, San Diego, 2005. `http://www.caida.org/publications/papers/2005/tr-2005-02/`.

[31] * B. Huffaker, M. Fomenkov, and k. claffy, “Internet Topology Data Comparison,” tech. rep., Center for Applied Internet Data Analysis (CAIDA), May 2012.

[32] * B. Huffaker, M. Fomenkov, and k. claffy, “Statistical implications of augmenting a BGP-inferred AS-level topology with traceroute-based inferences - Technical Report,” tech. rep., Center for Applied Internet Data Analysis (CAIDA), Nov 2016.

[33] Center for Applied Internet Data Analysis (CAIDA), “As rank.” `http://as-rank.caida.org/`.

[34] * 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.

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

[36] CAIDA, “CAIDA BGP Hackathon 2016.” `http://www.caida.org/workshops/bgp-hackathon/1602/`.

[37] B. Zhang, R. Liu, D. Massey, and L. Zhang, “Collecting the Internet AS-level Topology,” *ACM SIGCOMM CCR*, vol. 35, Jan. 2005.

[38] B. Augustin, B. Krishnamurthy, and W. Willinger, “IXPs: mapped?,” in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC ’09, pp. 336–349, 2009.

[39] Y. He, G. Siganos, M. Faloutsos, and S. Krishnamurthy, “Lord of the Links: A Framework for Discovering Missing Links in the Internet Topology,” *IEEE/ACM Transactions on Networking*, vol. 17, no. 2, pp. 391–404, 2009.

[40] A. Khan, T. Kwon, H.-c. Kim, and Y. Choi, “AS-level Topology Collection Through Looking Glass Servers,” in *IMC ’13*, 2013.

[41] V. Giotsas, S. Zhou, M. Luckie, and k. claffy, “Inferring Multilateral Peering,” in *CoNEXT ’13*, 2013.

[42] X. Shi, Y. Xiang, Z. Wang, X. Yin, and J. Wu, “Detecting Prefix Hijackings in the Internet with Argus,” in *IMC ’12*, 2012.

[43] V. Giotsas, “Periscope: tool and API,” 2016. `http://www.caida.org/tools/utilities/looking-glass-api/`.

[44] * V. Giotsas, A. Dhamdhere, and k. claffy, “Periscope: Unifying Looking Glass Querying,” in *Passive and Active Network Measurement Workshop (PAM)*, Mar 2016.

[45] * V. Giotsas, G. Smaragdakis, B. Huffaker, M. Luckie, and k. claffy, “Mapping Peering Interconnections to a Facility,” in *ACM SIGCOMM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, Dec 2015.

[46] K. Claffy, A. Dhamdhere, D. Clark, and S. Bauer, “First amended report of at&t independent measurement expert: Reporting requirements and measurement methods,” August 2016. `https://ecfsapi.fcc.gov/file/108042516812991/MB\%20Dkt\%2014-90\%20AT\&T\%20Inc.\%20First\%20Amended\%20IME\%20Report\%20ECFS.PDF`, also available at `https://www.caida.org/outreach/publications/`.

[47] * kc claffy, “Measuring internet interconnection performance metrics: an exercise to inform public policy,” February 2016. `http://www.caida.org/publications/presentations/2016/measuring_internet_interconnection_nanog/`.

[48] * Matthew Luckie and Amogh Dhamdhere and David Clark and Bradley Huffaker and kc claffy, “Challenges in Inferring Internet Interdomain Congestion,” in *ACM SIGCOMM Internet measurement Conference (IMC)*, 2014.

[49] * M. Luckie, A. Dhamdhere, B. Huffaker, D. Clark, and k. claffy, “bdrmap: Inference of Borders Between IP Networks,” in *Internet Measurement Conference (IMC)*, Nov 2016.

[50] S. Sundaresan, S. Burnett, N. Feamster, and W. De Donato, "BISmark: a testbed for deploying measurements and applications in broadband access networks," in *USENIX Annual Technical Conference (USENIX ATC 14)*, 2014.

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

[52] P. Ferguson and D. Senie, "Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing," May 2000. IETF BCP38.

[53] F. Baker and P. Savola, "Ingress filtering for multihomed networks," Mar. 2004. IETF BCP84.

[54] * E. Kenneally and K. Claffy, "Dialing Privacy and Utility: A Proposed Data-sharing Framework to Advance Internet Research," *IEEE Security and Privacy (S&P)*, July 2010.

[55] C. for Applied Internet Data Analysis, "UCSD Network Telescope," 2010. `http://www.caida.org/data/passive/network_telescope.xml`.

[56] A. Marder and J. M. Smith, "Map-it: Multipass accurate passive inferences from traceroute," in *Proceedings of the 2016 ACM on Internet Measurement Conference*, IMC '16, ACM, 2016.

[57] k. claffy, D. Clark, S. Bauer, and A. Dhamdhere, "Policy challenges in mapping Internet interdomain congestion," in *Telecommunications Policy Research Conference (TPRC)*, Oct 2016.

[58] "Measuring Broadband America.." `https://www.fcc.gov/general/measuring-broadband-america`.

[59] "Science Gateways Community Institute." `https://sciencegateways.org/`.

[60] "What is a Science Gateways: The Basics ." `https://sciencegateways.org/about/science-gateway-basics/`.

[61] B. Huffaker, "Interactive Access to Internet Topology Data," 2016. `http://www.caida.org/publications/presentations/2016/interactive_access_internet_topology_gateways/`.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[76] 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*, vol. 34, no. 7, 2016.

[77] U. Pisa, "NTOP traffic monitoring software." `ntop.org`.

[78] "CAIDA Annual Report," 2015. `https://www.caida.org/home/about/annualreports/2015/#data`.

[79] Center for Applied Internet Data Analysis (CAIDA), "Papers Published (by non-CAIDA Authors Using CAIDA Datasets." http://www.caida.org/data/publications/.

[80] "CAIDA papers." http://www.caida.org/publications/.

[81] G. Nomikos and X. A. Dimitropoulos, "traIXroute: Detecting IXPs in traceroute paths," in *PAM*, 2016.

[82] I. Cunha, R. Teixeira, D. Veitch, and C. Diot, "Predicting and tracking Internet path changes," in *ACM SIGCOMM*, 2011.

[83] U. Javed, I. Cunha, D. R. Choffnes, E. Katz-Bassett, T. Anderson, and A. Krishnamurthy, "PoiRoot: Investigating the root cause of interdomain path changes," in *ACM SIGCOMM*, August 2013.

[84] E. Katz-Bassett, C. Scott, D. R. Choffnes, I. Cunha, V. Valancius, N. Feamster, H. V. Madhyastha, T. E. Anderson, and A. Krishnamurthy, "LIFEGUARD: practical repair of persistent route failures," in *ACM SIGCOMM*, 2012.

[85] Y. Shavitt and N. Zilberman, "Improving IP Geolocation by Crawling the Internet PoP Level Graph," in *Networking*, 2013.

[86] Y. Shavitt and N. Zilberman, "A Structural Approach for PoP Geolocation," *Proceedings of the 2010 IEEE INFOCOM Conference*, 2010.

[87] A. H. Rasti, N. Magharei, R. Rejaie, and W. Willinger, "Eyeball ASes: from geography to connectivity," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2010.

[88] Z. Hu and J. Heidemann, "Towards Geolocation of Millions of IP Addresses," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2012.

[89] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards street-level client-independent IP geolocation," in *USENIX Symposium on Networked Systems Design & Implementation (NSDI)*, 2011.

[90] * B. Huffaker, M. Fomenkov, and kc claffy, "Geocompare: a comparison of public and commercial geolocation databases," tech. rep., Center for Applied Internet Data Analysis, 2011. `http://www.caida.org/publications/papers/2011/geocompare-tr/`.

[91] * I. Livadariu, A. Elmokashfi, A. Dhamdhere, and kc claffy, "A first look at IPv4 transfer markets," in *CoNEXT*, 2013.

[92] * Alberto Dainotti and Phillipa Gill, "NSF CNS-1423659. HIJACKS: Detecting and Characterizing Internet Traffic Interception based on BGP Hijacking," 2014. `http://www.caida.org/funding/hijacks/`.

[93] * CAIDA, "AS Classification: method for classifying ASes according to business type," 2016. `http://www.caida.org/data/as-classification/`.

[94] G. Huston, "The death of transit," October 2016. `https://blog.apnic.net/2016/10/28/the-death-of-transit/`.

[95] Mark Jamison, "Do We Need the FCC?," *Tech Policy Daily*, December 2016. `http://www.techpolicydaily.com/communications/do-we-need-the-fcc/`.

[96] W. Lehr, E. Kenneally, and S. Bauer, "The Road to an Open Internet is Paved with Pragmatic Disclosure and Transparency Policies," in *Telecommunications Policy Research Conference (TPRC)*, Sep 2015.

[97] k. claffy, D. Clark, S. Bauer, and A. Dhamdhere, "Policy challenges in mapping Internet interdomain congestion," in *Telecommunications Policy Research Conference (TPRC)*, Oct 2016.

[98] Rob Frieden, "Case Studies in Abandoned Empiricism and the Lack of Peer Review at the Federal Communications Commission," August 2009. `https://ssrn.com/abstract=1456516`.

[99] S. Elaluf-Calderwood and J. Liebenau, "Idea to Retire: Internet without policy metrics," March 2016. `https://www.brookings.edu/blog/techtank/2016/03/02/idea-to-retire-internet-without-policy-metrics/`.