Project Summary

Internet cartography has emerged as a new field of computer as well as network science, with several global Internet measurement infrastructures executing comprehensive topology mapping measurement experiments, continuously, for years. UCSD's Center for Applied Internet Data Analysis (CAIDA) has operated the longest-running of these measurement infrastructure platforms (Archipelago), which has supported scientific measurement experiments of the global Internet since September 2007. This platform has collected 90 billion traceroutes in 39 TB of files, growing 16 billion traces and 7 TB annually (5-year doubling rate). These data sets have already yielded impacts across a broad range of CISE sub-disciplines. Yet the biggest remaining obstacle to even more productive scientific use of this unbounded wealth of information is infrastructural: the lack of an easy-to-use and analytically powerful exploratory interface to the data.

Researchers have made explicit requests for search functionality that would have a transformative impact on several focused areas of CISE-funded research. In response to community feedback, we propose to develop the FANTAIL system – Facilitating Advances in Network Topology Analysis – to enable discovery of the full potential value of massive raw Internet end-to-end path measurement data sets. We envision a four-component system: (1) an interactive web interface; (2) an API built on web standards; (3) a full-text search system based on Elasticsearch; and (4) a big data processing system based on Spark, leveraging SDSC's cluster resources. Although our goal is to enhance the general accessibility and utility of this data, our project will be driven by specific compelling use cases, in response to research community needs for interactive exploratory capabilities. To this end, we will identify and implement reusable components, analysis modules, which will serve as primitives for constructing more complex data-processing pipelines. Users will specify, via the web interface or API, a sequence of analysis modules to execute on the set of traceroute paths matched by their queries. The FANTAIL system will then perform the queries, run the analysis modules, and provide the output for download for further analysis or processing by researchers on their own systems. We will implement analysis modules that are useful for (1) performing data reduction (to minimize the amount of data users have to download and process), (2) enhancing raw traceroute data with various annotations available publicly or created by us, and (3) offloading commonly-needed analysis/data processing tasks from users.

Intellectual Merit

A large-scale topology query system for searching CAIDAs and other large traceroute data archives will enable a broad set of networking and security research areas in CISE that rely on the emerging sub-discipline of Internet cartography. Based on community feedback, we believe the proposed system will enable fundamentally new research directions for researchers who have applied CAIDA's topology data over the last two decades to a variety of areas related to our understanding of the Internet as critical infrastructure.

Broader Impacts

The results of our project will greatly improve the research utility of the accumulated huge archives of Internet topology and performance data and simplify its use. We will widely advertise the platform availability and new developments via project web pages, community meetings, proposed workshops, and CAIDA blog. Our proposed solutions will lower barriers for entering into the field of network science and computing research thus leading to increased diversity of our user base. We will also apply for REU funds in order to support training a new generation of engineers, programmers, and Internet researchers.

Contents

Project Description: CCRI: Medium: FANTAIL - Facilitating Advances in Network Topology Analysis

1 Introduction, Motivation and Goals

Internet cartography has emerged as a new field of computer as well as network science, with several global Internet measurement infrastructures executing comprehensive topology mapping measurement experiments, continuously, for years. UCSD's Center for Applied Internet Data Analysis (CAIDA) has operated the longest-running of these measurement infrastructure platforms (Archipelago) [1, 2], which has supported scientific measurement experiments on the global Internet from vantage points (mostly Raspberry Pi devices) hosted in academic, commercial, and residential networks around the globe since September 2007. Since then, this platform has collected 90 billion traceroutes in 39 TB of files, growing 16 billion traces and 7 TB annually (5-year doubling rate). Europe's regional Internet address registry RIPE has also deployed a global network of probes (RIPE Atlas), primarily to support operational assessments of reachability. In 2016, RIPE's platform began to more comprehensively collect path information targeting the whole IP address space, similar to CAIDA's probing [3]. Google's Measurement Lab (M-Lab) was also collecting 7 million traceroutes per month in May 2017, after Google launched a new policy to trigger an NDT measurement (which includes a traceroute) to a sample of IP addresses that visit the Google search page [4]. These billions of Internet path measurements, terabytes of data, from thousands of vantage points supported by all of these platforms, sit in storage facilities around the world, constituting a largely untapped data source about the structure and dynamics of the most important complex network ever "designed". The last decade has yielded research breakthroughs that have overcome many methodological obstacles to using this data: dealing with artifacts common in traceroute data [5, 6], mapping different IP addresses to the same router [7, 8, 9, 10, 11, 12, 13]; and mapping IP addresses to ASes and boundaries between ASes [14, 15]. The biggest remaining obstacle to scientific use of this unbounded wealth of information is infrastructural: the lack of an easy-to-use and analytically powerful exploratory interface to the data.

In response to community feedback, we seek to develop the FANTAIL system – Facilitating Advances in Network Topology Analysis – to enable discovery of the full potential value of massive raw Internet end-to-end path measurement data sets. We propose a four-component system: (1) an interactive web interface; (2) an API built on web standards; (3) a full-text search system based on Elasticsearch [16]; and (4) a big data processing system based on Spark [17], leveraging SDSC's cluster resources [18]. Although our goal is to enhance the general accessibility and utility of this data, our infrastructure development agenda will be driven by specific compelling use cases, in response to the research community requests for interactive exploratory capabilities, ideally ten seconds or less per query over billions of traceroutes. To this end, we will identify and implement reusable components, *analysis modules*, that we will execute on behalf of, and under the direction of, users. These analysis modules will serve as basic primitives for constructing more complex data-processing pipelines. Users will be able to specify, via the web interface or API, a sequence of computations to apply to the set of traceroute paths matched by their queries. The FANTAIL system will then perform the queries, run the analysis modules, and provide the output for download for further analysis or processing by researchers on their own systems, if needed. We will implement analysis modules that are useful for (1) performing data reduction (to minimize the amount of data users have to download and process), (2) enhancing raw traceroute data with various annotations available publicly or created by us, and (3) offloading commonly-needed analysis/data processing tasks from users.

This proposal is structured in specific sections required by the CCRI solicitation. Section 2 describes the relationship of the proposed infrastructure to existing platforms, data, and technologies. Section 3 describes the proposed infrastructure, including: the querying and big data processing components; ancillary tools, resources, and data sets; user services; community engagement; and outreach. Section 4 provides details on how this research infrastructure will support CISE researchers pursuing a variety of focused research agendas that all rely on Internet cartography, itself an emerging sub-discipline of computer and information science and engineering. The remaining sections cover sustainability, intellectual merit, broader impacts, management plan, and results from prior CRI support.

2 Relationship of proposed infrastructure to existing resources

In addition to consultations at annual workshops and one-on-one interviews with a diverse set of users of traceroute data (Section 4), our long history of efforts in making traceroute data more useful to the community has provided us with extensive preliminary community input and technical experience that drive this proposed project. We review the most important of these efforts here.

2.1 Henya

Our previous work, Henya [19], is the only known public system for searching large archives of Internet path measurement data. This prototype search system consists of an interactive web interface along with a traceroute search engine. Henya allows users to conduct some of the queries on general traceroute properties listed in Table 1 and all of the queries on topological properties listed in Table 2. In addition, the Henya web interface provides a built-in analysis and visualization of the RTT time series for target destinations, an RTT histogram, and a time series of target unreachability. The Henya search engine is implemented in the scripting language Python [20] and uses the RocksDB [21] key-value data store for traceroute and custom indexing data.

We have indexed 21.5 billion traceroute paths (8.8 TB of traceroute data) in Henya, a subset of CAIDA's comprehensive, ongoing topology datasets.

Henya is able to provide fast queries of large volumes of traceroute data because of its novel indexing data structures that exploit the unique properties of traceroute data and the append-only nature of large traceroute datasets. Henya achieves its performance by preprocessing and analyzing traceroute data offline and then generating several specially-designed indexing data structures to be used online during querying (the immutability of the data allows us to do this offline processing). The most important Henya indexing data structure records the occurrences of IP addresses in multiple time scale bins, which allows rapid narrowing of the search to relevant traceroute paths when executing complex queries about neighboring addresses (see Sec. 3.1.1).

The FANTAIL system builds on our knowledge and experience gained with Henya and will be Henya's *spiritual* successor but with a much broader scope and scale. FANTAIL will couple Henya-like searching abilities with a big data processing component, which will enable a wide variety of analyses that were not possible with Henya alone. In addition, despite Henya's successful demonstration of searching at large scale, we will start afresh with the open-source Elasticsearch [16] full-text search engine. We concluded it would be a better use of our time and effort to switch to an off-the-shelf full-text search engine than to rewrite the Henya search engine in a compiled language like C or C++ (Python is up to 100x slower than C/C++) and to add support for multicore and cluster computing to the Henya engine (Elasticsearch has this support). As necessary, we will apply the novel indexing structures and algorithms of Henya in the design and

traceroute to 200.136.34.2 (sao2-br.ark.caida.org) from **bjc-us** of *commercial network (6)* using ICMP

Hop	Address	Prefix	AS	Location	RTT (ms)
1	**unknown.Level3.net** 209.245.28.1	209.244.0.0/14	3356	broomfield, co usa	0.3
2	**ge-5-0-48.hsa2.Denver1.Level3.net** 209.245.29.226	209.244.0.0/14	3356	denver, co usa	0.8
3	**ge-7-36.car2.Denver1.Level3.net** 4.69.200.66	4.0.0.0/9	3356	denver, co usa	1.9
4	**vlan51.ebr1.Denver1.Level3.net** 4.69.147.94	4.0.0.0/9	3356	denver, co usa	0.8
5	**ae-2-2.ebr2.Dallas1.Level3.net** 4.69.132.106	4.0.0.0/9	3356	dallas, tx usa	15.0
6	**ae-72-72.csw2.Dallas1.Level3.net** 4.69.151.141	4.0.0.0/9	3356	dallas, tx usa	15.0
7	**ae-2-70.edge2.Dallas1.Level3.net** 4.69.145.75	4.0.0.0/9	3356	dallas, tx usa	15.6
8	**DATA-RETURN.edge2.Dallas1.Level3.net** 4.71.220.70	4.0.0.0/9	3356	dallas, tx usa	15.1
9	**g1-10.br1.dfw.terremark.net** 66.165.160.249	66.165.160.0/19	23148	dallas, tx usa	47.1

Figure 1: Sample traceroute path annotated with DNS hostname, BGP prefix, and autonomous system (AS)

implementation of indexing on top of Elasticsearch to achieve the massive increase in scalability envisioned with FANTAIL.

2.2 CAIDA Ark and RIPE Atlas topology measurements

The proposed infrastructure will make use of a number of datasets produced by CAIDA's (previously CRI-funded) active measurement infrastructure Archipelago (Ark) [1, 2], which consists of around 200 measurement nodes (vantage points) where vetted researchers can run experiments [22, 23, 13, 9, 10, 24, 6, 25, 26].

The FANTAIL system will provide access to the following comprehensive, ongoing Internet topology data sets produced by Ark: the *IPv4 Routed /24 Toplogy Dataset* [27] and the *IPv4 Prefix-Probing Traceroute Dataset* [28]. The Routed /24 dataset systematically gathers IP-level paths to a dynamically generated list of IP addresses covering each /24 network in the routed IPv4 space. The Prefix-Probing dataset probes each routed prefix from each monitor once per day, to allow the analysis of temporally finer-grained dynamics, e.g., route hijacking. Since 2007 we have gathered over 90 billion traceroutes (over 39 TB) in these two datasets. We project these datasets to double in size within 5 years at current data collection rates, or sooner if we significantly increase the number of measurement nodes.

The RIPE Atlas platform consists of over 10,000 probes, which conduct a variety of measurements, including DNS lookups, SSL certificate downloads, HTTP downloads, and traceroute measurements. Traceroute measurements are either built-in or user-defined, the latter conducted by Atlas users in the community. We will initially make the most recent few years of Atlas traceroute measurements available in FANTAIL, which will enable rigorous comparisons of the relative coverage of these two infrastructures and their epistemological implications (Section 4.1).

2.3 Derivative data sets to support methodologically sound use of data

Some research questions do not require the raw data, and interpretation of raw traceroute data is methodologically challenging due to a variety of artifacts that may appear in reported paths. Rather than every researcher having to repeatedly tackle these challenges, based on community feedback years ago, we began to regularly curate and share derivative data sets for researchers

[29, 30, 31, 32]. Additionally, we distill two-week snapshots of raw traceroute data into Internet Topology Data Kits (ITDKs) [33] and infer annotated router-level and AS-level topologies of the global Internet. We have increased the richness of these annotations over time by integrating new techniques as we have developed them: mapping IP addresses to routers (using CAIDA's MIDAR, iffinder, and kapar techniques [12]), required to construct physical router-level topologies [13]; mapping routers to the autonomous systems (ASes) that own them using CAIDA's latest *bdrmapIT* tool [15] and extracting network boundaries from raw traceroute data (also using the bdrmapIT tool [15]). We also perform DNS lookups of all IP addresses found in our traceroute data [34]. DNS hostnames often encodes topologically relevant information, e.g., link type (backbone vs. access), link capacity, Point of Presence (PoP), and geographic location. We will integrate these auxiliary data into FANTAIL as annotations on traceroute paths or as inputs to built-in big data analyses (see Fig. 1 for a sample traceroute path annotated with various auxiliary information).

These data sets, which we propose to make more accessible have already yielded impacts across a broad range of CISE sub-disciplines (Section 4). But researchers have made explicit requests (see LOCs) for search functionality that would have a transformative impact on several focused areas of CISE-funded research (Section 4).

3 Infrastructure Description

This section describes the proposed infrastructure, including: the traceroute querying and big data processing components; ancillary tools, resources, and data sets; user services; community engagement; and outreach.

3.1 Fundamental infrastructure

The proposed system consists of the following four main components: (1) an interactive web interface, (2) an API built on web standards, (3) a full-text search system based on Elasticsearch [16], and (4) a big data processing system based on Spark [17]. We will use SDSC cluster resources [18] for running Elasticsearch and Spark (resources committed by SDSC, see LOC).

The central data type of the system is the *traceroute path*, which represents the inferred IP-level Internet path that network traffic would take between two hosts, the measurement *vantage point* and the *destination*, as determined with the traceroute technique by a variety of tools, such as traceroute [35] and scamper [36] (CAIDA uses scamper to collect our traceroute datasets as well as for other measurement experiments).

The system will operate as follows:

1. Accept user requests via the web interface/API.
2. Execute queries to find traceroute paths of interest.
3. Optionally execute big data processing on matching traceroute paths.
4. Provide, for download by the user, the matching traceroute paths or the output of big data processing.

3.1.1 Traceroute path querying component

For the purposes of querying, a traceroute path consists of: the vantage point address; the destination address; an ordered sequence of traceroute hops with IP address and round-trip time (RTT); path length; whether we reached the destination; and whether there is observable evidence of MPLS tunnels in the path [37]. MPLS tunnels are used for traffic engineering, but they represent hidden topology to traceroute exploration, and failure to account for them yields incomplete and inaccurate topology inferences [38].

Table 1: *Summary of queries matching general traceroute properties. Vantage point organization types include residential, academic, business, etc. An operator (op) suffix to a query name can be* `lt`, `eq`, *or* `gt` *to mean* $<$, $=$, *or* $>$ *respectively.*

Query	Selection Criteria
`vp` V	vantage point is V
`vp_as` N	vantage point is located in autonomous system (AS) N
`vp_country` C	vantage point is located in country C
`vp_type` T	vantage point is hosted by an organization of type T
`status` N	traceroute has success/failure code N
`timestamp_`*op* N	traceroute has timestamp $<$, $=$, or $>$ N
`dest_rtt_`*op* N	RTT of traceroute destination is $<$, $=$, or $>$ N ms
`pathlen_`*op* N	length of traceroute path is $<$, $=$, or $>$ N
`has_mpls` T/F	whether there is (T) or is not (F) MPLS in the traceroute path

Table 2: *Summary of types of queries matching IP addresses in the traceroute path. The parameter T represents a single target address or network prefix. The parameter G represents a target group which is a list of one or more targets—that is, $G = T_1, \ldots, T_n$. A target group specifies the union of a set of targets—that is, $G = T_1 \cup \ldots \cup T_n$.*

Query	Selection Criteria
`dest` T	traceroute destination is any address $t \in T$
`hop` T	traceroutes with any address $t \in T$ appearing at any hop
`neigh` $G_1 \ldots G_n$	traceroutes with n distinct neighboring hop addresses $t_i \in G_i$

Based on feedback from researchers at our workshops (some of which is reflected in the letters of collaboration), we propose to provide a number of query primitives for searching traceroute paths in various ways; these primitives will support the construction of complex queries by the combination of any number of query primitives with the logical AND and OR operators. Table 1 lists the queries that match on various general properties of traceroutes, including the vantage point, traceroute collection timestamp, success/failure status, path length, end-to-end round trip time, and the presence of MPLS tunnels on the path.

Table 2 lists the queries that match on the IP addresses of the traceroute path. The parameter T represents a single *target* address or network prefix. The parameter G represents a *target group* which is a list of one or more targets—that is, $G = T_1, \ldots, T_n$. A target group specifies the union of a set of targets—that is, $G = T_1 \cup \ldots \cup T_n$. For example, `1.0.0.0, 2.0.0.0, 3.0.0.0/16`[1] is a target group consisting of two IP addresses and one prefix, and an address t would match this group if $t =$ `1.0.0.0` or $t =$ `2.0.0.0` or $t \in$ `3.0.0.0/16`. Among other things, target groups are useful for specifying (1) *router aliases*—that is, the IP addresses assigned to the network interfaces of a single router as discovered by, for example, the MIDAR alias resolution tool [13], (2) the prefixes announced by an autonomous system (AS), (3) the prefixes announced by ASes in a given country, and (4) the prefixes used at an Internet Exchange Point (IXP) for peering.

The `dest` T query is useful for finding traceroutes conducted toward a given IP, prefix, AS, or country. The `hop` T query is useful for finding traceroutes that cross a given interconnection link (taking aliases into account, if alias data is available), AS, country, or IXP.

[1]IP address prefixes are written in CIDR notation: the suffix indicates the number of bits in the prefix; the remaining bits are available to address hosts within the prefix.

The `neigh` $G_1 \ldots G_n$ query is the most powerful – it provides a way to find traceroute paths that contain two or more **neighboring** hops, which may appear in any order and be separated by any number of hops, or be constrained in their order and separation distance. As an example, suppose we performed the query `neigh 1.0.0.0/24 2.0.0.0/24`. This matches traceroutes that meet the following two conditions: there is a hop with an address $t_1 \in$ `1.0.0.0/24` **and** a (different) hop with an address $t_2 \in$ `2.0.0.0/24`. Furthermore, if these two target prefixes represent two different ASes (in reality, most ASes announce more than one prefix), then the matching traceroutes will provide all observed paths between the ASes. This ability to query for neighboring hops is essential for gathering data that can allow a researcher to study, for example, the interconnection links, IXPs, and transit providers used between two autonomous systems (ASes).

The traceroute queries described in Tables 1 and 2 represent high-level query primitives expressed in the language of traceroute data. FANTAIL users will work with these queries. We will implement code to map these high-level queries into low-level calls to the Elasticsearch search API. We will also transform or augment the traceroute data as needed to enable efficient indexing and execution of these types of queries under the inverted-index model of full-text search provided by Elasticsearch.

3.1.2 Big-data processing component

Making the most of a big data processing system requires a non-trivial investment in time and effort to gain the necessary knowledge and skills to effectively develop the data processing code to run on such a system. We cannot expect the potential users of our system to be willing (and/or able) to incur such a cost – our goal is to lower the barrier to using big data processing as much as possible in order to provide access to the widest audience of researchers as possible, including to researchers outside traditional network research domains.

To this end, rather than providing direct access to the big data processing backend, we will identify and implement reusable components, *analysis modules*, that we will execute on behalf of, and under the direction of, users. Analysis modules serve as basic primitives for constructing more complex data-processing pipelines. Users will be able to specify, via the web interface or API, a sequence of analysis modules to execute on the set of traceroute paths matched by their queries. The FANTAIL system will then execute the queries and the analysis modules and provide the output for download for further analysis or processing by the researchers on their own systems, if needed.

We will implement analysis modules that are useful for (1) performing data reduction (to minimize the amount of data users have to download and process), (2) enhancing raw traceroute data with various annotations available publicly or created by us, and (3) offloading commonly-needed analysis/data processing tasks from users so that users do not need to reinvent the wheel.

The following are the initial set of analysis modules we propose to implement:

- Reduce a set of traceroute paths that match a given property (see Tables 1 and 2) to the set of unique paths or to a graph, where nodes are IP addresses and links are observed adjacencies between IP addresses in paths.
- Annotate IP addresses with IP aliases discovered in CAIDA's Internet Topology Data Kits (ITDKs) (Section 2.3), to enable router-level topology mapping (see Waikato, UPenn LOCs).
- Identify IP addresses belonging to Internet Exchange Points (IXPs) using publicly available information on which BGP prefixes belong to IXPs [39] (see FORTH LOC, a CISE-funded collaboration with CAIDA).
- Map IP addresses in traceroute paths to network operators (ASes) using bdrmapIT [14, 15, 40] data generated from CAIDA's ITDKs (see UPenn LOC, a CISE-funded collaboration).

- Identify interconnection points in traceroute paths using bdrmapIT data [15] (see UPenn LOC, a CISE-funded collaboration).
- Identify MPLS tunnels using TNT measurements [38] (see U. Liege LOC, a CISE-funded collaboration with CAIDA researchers).[2]
- Extract, analyze, and compute various statistics on round-trip time data; e.g., round-trip time series from a given vantage point to a set of destinations.

In addition to analysis modules, we will implement *analysis recipes*, which are potentially complex multi-step querying and processing pipelines that we will provide prepackaged for users to execute as a unit with a single click. Users will customize the execution of analysis recipes by supplying parameter values, such as the target AS.

The following are the initial set of analysis recipes we propose to implement:

- *Find all interconnection links of a given AS or set of ASes (sibling ASes) in a given time period.* Implementation: (1) Find all traceroutes that have IP addresses in the target ASes, (2) identify interconnection links in the selected traceroutes with bdrmapIT data, and (3) produce the unique set of interconnection links from these traceroutes.
- *Construct the topology of a target AS (that is, all routers and links that map to the AS) in a given time period, presented as a graph.* Implementation: (1) Find all traceroutes that have addresses in the target AS, (2) identify interconnection links with bdrmapIT data to determine the exact boundary of the AS, (3) generate an IP-level graph from the traces that only includes interfaces and links in the target AS, and (4) apply ITDK alias-resolution data to overly a router-level graph on the IP-level graph.

We will use project workshops to engage with the community to identify other reusable analysis modules and recipes that are worth implementing.

3.2 Tools, resources, and data sets

We will develop tools to import CAIDA's and RIPE's traceroute data into Elasticsearch, both historical data and new data as they become available (we will import the full archive of CAIDA data and the most recent few years of RIPE's). For RIPE, this will involve setting up regular downloading of new traceroute data from the RIPE Atlas data web site. Additionally, Elasticsearch requires indexed content to be in JSON format with a common document structure. RIPE's data is already in JSON format, but CAIDA's data is in the binary scamper *warts* format. There is a tool available in the scamper package for converting warts to JSON, but the resulting JSON document structure differs significantly from RIPE's. We will develop tools to convert these two JSON formats into a common structure and to transform/augment the traceroute data in whatever ways are needed to allow Elasticsearch to efficiently index the data; e.g., for each traceroute path, compute the path length and check for the presence of MPLS ICMP extensions, and then directly store these values in new fields in the JSON document for ease of indexing.

A number of analysis modules described in Sec. 3.1.2 involve annotating traceroute paths with additional data. We will develop tools to assemble and store these auxiliary data into databases for use by our system. Auxiliary data include: DNS lookups of all observed IP addresses in CAIDA traceroute paths; list of network prefixes used at IXPs; bdrmapIT and alias-resolution data in ITDK releases; and MPLS TNT data. Many of these auxiliary datasets are produced by CAIDA on the Ark platform.

Consistent with the CCRI solicitation, we have also considered a list of possible future enhancements, as resources allow, or other researchers are inspired to contribute:

[2] We have about 4 months of continuously collected TNT data from 28 Ark vantage points.

1. Provide the results of traceroute queries, analysis modules, and analysis recipes in a dashboard or visualization (e.g., RTT distributions, topology visualizations).
2. Implement additional analysis modules/recipes, such as ones to support investigations of grey market IP address transfers, e.g., find all ingress links to the target prefix/AS in a given time period; or extract a time series of ingress links to the target prefix/AS, and correlate with anomalous changes in DNS and BGP data [41].
3. Index IPv6 traceroutes from CAIDA and RIPE.

3.3 User services

We will provide two complementary ways of accessing the infrastructure: an *interactive web site* and an *API*. The *web site* will allow researchers to perform traceroute queries; construct and execute a data processing pipeline by selecting and arranging available analysis modules; and execute analysis recipes, filling out parameters via a web form. Researchers will be able to bulk download results in a suitable format (JSON, CSV, etc.) via their web browser. For interactive traceroute queries returning a small number of results, the researcher will be able to view the traceroute paths annotated with auxiliary data, including DNS, router ownership, layer 2 information, IP aliases, exchange facility presence, interconnection links, and BGP prefix, depending on availability of data (not all auxiliary data is available for the full time period of our data collection). The *query API* will provide the same features of the web site, but use an API built on web standards (HTTP, JSON, etc.).

To make the infrastructure readily available to researchers, we will advertise FANTAIL platform, and announce its subsequent improvements/expansions, on the CAIDA web site. To request access, researchers will have to fill in a form online and electronically sign the Acceptable Use Agreement (AUA). CAIDA Community Liaison personnel will review each application and grant access based on the merits of the proposed research use. This authorization and authentication approach leverages CAIDA's existing and proven data sharing mechanisms [42] while also allowing us to vet prospective users and annually review their needs and activities.

3.4 Community Engagement

Several CISE researchers will be engaged in the design, development, and management of the infrastructure, detailed in Section 4 and letters of collaboration. In particular, all collaborators have agreed to participate in our workshops, and provide input on additional functionality to support research community needs. We will conduct annual written user surveys to evaluate user satisfaction, as well as group survey discussion at the workshops. We will participate in CCRI Virtual Organization, and CCRI community PI meetings, as well as present results of this project to academic and operational networking and security communities, e.g., NANOG, as relevant. We will also foster community use of this new platform through the DHS-funded Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT) project, which will allow us to leverage cross-agency support for Internet infrastructure data sharing activities.

3.5 Community Outreach

Our two annual workshop series – AIMS in its 10th year and WIE its 9th – have established solid communities of diverse collaborators in technical and social science fields, respectively. The Active Internet Measurement Systems workshop series 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 [43, 44, 45, 46, 47, 48, 49, 50, 51]. 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 [52, 53, 54, 55, 56, 57, 58, 59, 60], including policy-relevant research funded by CISE (see UCI LOC). We will use these workshop series to build and nuture a robust and diverse community of FANTAIL users: introduce new capabilities including hands-on tutorials to engage users; soliciting feedback on user satisfaction to improve subsequent infrastructure interfaces and operation; inspire the use of FANTAIL in multi-disciplinary collaborations; and share experiences with classroom use of FANTAIL.

We will open a project web site highlighting the new capabilities available; a wiki will provide orientation material and documentation for all the tools and datasets indexed in the platform. We will create a mailing list of FANTAIL users to facilitate circulation of ideas and promote the creation of a community. This channel for information exchange will be especially useful if hands-on experience with real-time data analysis will prompt researchers to contribute to improving available tools for monitoring, analysis, and data visualization of the telescope data.

4 CISE research opportunities enabled by proposed infrastructure

We have taken two steps to identify research opportunities enabled by the proposed infrastructure: consultations at our AIMS measurement workshops, and one-on-one interviews with collaborators and colleagues. Based on these conversations, in this section we summarize the research opportunities to be enabled by the infrastructure. Table 3 lists a sample of research interests articulated in the attached letters of collaboration, which reflect a diverse community of users.

Table 3: *CISE-funded research topics described in letters of collaboration (LOCs).*

Enabled Research	Letters of Collaboration
Scientific Network Modeling & Mapping	
· *inferring accurate maps:*	
· · · alias resolution	U. Liege, Columbia
· · · interconnection modeling (economics)	MIT, UC Irvine
· · · device fingerprinting	U. Liege
· · · inference of layer-2 (MPLS) topology	U. Liege, U Penn.
· · · routing and topology, IXP analysis	Columbia, MIT
· *longitudinal studies*	MIT
· IPv6	NPS, Liege
Security and Stability Vulnerabilities	
· outages	CAIDA, NPS, U.MD
· congestion	CAIDA, MIT
· attack detection	Parsons, CAIDA
· hijacking/routing security	CAIDA, Parsons
· infrastructure risk assessments	Parsons, NPS, Waikato, IIJ, CAIDA
· middleboxes	Liege

4.1 Networking Research

The networking research community, including a broad set of sub-disciplines in CISE, has relied on CAIDA's measurement and data infrastructure for decades, as evidenced by our statistics on data usage (reported annually e.g., [61]) and papers published using our data [62] (including our own [63]). We focus on several specific CISE-funded networking research areas that will be

enabled by the proposed infrastructure development: layer-2 topology mapping; device fingerprinting; routing research; and interconnection structure and dynamics.

CAIDA has collaborated with Benoit Donnett (U. Liege, see LOC) for years to develop and evaluate measurement methods and implementations that reveal **layer-2 network topology**, most notably MPLS tunnels that can reduce the completeness and accuracy of IP topology inference. PI Claffy recently served on the PhD dissertation committee of Yves Vanaubel, who spent 6 months at CAIDA to use the Archipelago platform to deploy and test the most state-of-the-art methods for MPLS tunnel detection [38]. His method extends Paris traceroute to reveal most (if not all) MPLS tunnels along a path. First, using traceroute probes, his tool looks for evidence of hidden tunnels. e.g., abrupt and significant TTL shifts in traceroute output. If detected, the tool launches additional dedicated probing to discover details of the hidden tunnel. This tool has been deployed on the Archipelago platform for 5 months. We will use FANTAIL to continue to collaborate with Donnet to investigate the impact of MPLS tunnels on topological analysis, as well as analyze trends in MPLS tunnel infrastructure evolution over time.

Furthermore, this technology enables improved **network fingerprinting** (Prof. Donnett currently has a masters student working on this topic), notably for being able to identify Cisco, Juniper, or Huawei devices. The reality is that enterprises, including campuses, often do not have a complete inventory of equipment deployed by vendor, so this technology would allow discovery of otherwise unknown and unwanted devices on their networks.

Vendor-specific device fingerprinting can also be used to improve methods for **IP address alias resolution**, because it will narrow the set of candidate IP interface addresses that could be on the same device [64]. That is, if two IP interfaces are candidates for alias resolution, yet both IP differ on their respective fingerprinting, we can immediately discard those candidates.

Ethan Katz Bassett (see LOC Columbia) has undertaken **interdomain routing research** for over a decade, including developing a "practical Internet route oracle" system that takes queries expressed as regular expressions and returns traceroutes that match, even if the system has never measured a matching path [65]. He presented this idea at an early AIMS workshop, soliciting feedback before building his system. This system tried to support queries over measurements that had not been issued yet, in contrast to FANTAIL which will support queries over historical measurements. Systems such as his could use historical queries from FANTAIL to help improve its decisions of which measurements to issue. Katz-Bassett has requested the capability to search traceroute archives for specific IXPs, prefixes, and across specific times, as it would enable a source of validation for his research that is not currently available. Several faculty have indicated they would use such a capability for teaching students about the complexity of the Internet topology.

Two of our collaborators are focused on interconnection, both the structure (LOC UPenn), as well as dynamics (both traffic and economic, LOC UC Irvine and LOC MIT.) Jonathan Smith is the PhD advisor of Alex Marder, developer of the MAPIT algorithm and collaborator with CAIDA on the bordermapIT system that extracts network boundaries from traceroute archives [15]. FANTAIL will provide the unprecedented opportunity to search topology data annotated with infrastructure ownership inferences using these state-of-the-art methods.

FANTAIL will also support study of **performance problems**, including **persistent interdomain congestion** (LOC MIT) due to peering disputes. Researchers can use FANTAIL to identify source-destination pairs that traverse a given interdomain link, and probe these destinations to discover evidence of congestion along the path [66, 67, 68] (MIT and UPenn LOCs).

In another CISE-funded NeTS project (see LOC), UCI Professor Scott Jordan is developing methods and metrics to study Internet interconnection. Interconnection disputes have led to failures to augment interconnection capacity, which in turn have harmed the experience of consumers. Dr. Jordan is developing a model of Internet interconnection that incorporates technical

and economic factors. He has specifically asked for functionality to search large archives of traceroute data for specific interconnection points, IP prefixes, geographic information, and across specific times, to enable an understanding of which ISPs interconnect with which transit providers and content delivery networks, and where. With the proposed tool, we will be able to analyze massive archives of IP level paths to refine his picture of interconnection structure and dynamics.

Finally, there is a growing need to understand how to most efficiently deploy global Internet measurement infrastructure. The RIPE platform consists of several thousand probes, but only 450 are *anchors* hosted in reliable networks with known locations. The CAIDA Ark platform has over 200 monitors also with known locations. Each platform performs regular traceroute measurements with differing frequencies and targets, but no one has quantified the overlap or lack thereof and consequent implications for research use of these data. FANTAIL will facilitate this type of analysis. We will initially index several years the RIPE Atlas traces to enable geographic and topological analysis of the divergence of the resulting measurements, and consider implications for reproducibility and validation of scientific inferences using the measurements (see RIPE LOC). This effort will improve our understanding of network visibility, but enable more efficient of investment in future Internet measurement infrastructure, a primary focus of UIUC's CISE-funded PacketLab project (CNS 1763884 and 1903612, with a subcontract to UCSD).

4.2 Security and Stability research

There are also CISE-funded projects in Internet infrastructure security and stability research: outages, infrastructure risk assessments, attacks, route hijacks – that will benefit from flexible and convenient access to the wealth of topology data. An increasingly important area is the susceptibility of different parts of the topology, e.g., countries, to disruption.

Several research groups, in addition to CAIDA researchers, are improving and extending methods for **detecting and quantifying the impact of wide-area Internet outages** by combining active and passive measurements [69, 70, 71, 72, 73, 74, 75]. A related body of recent research focuses on **predicting, tracking, and localizing the root cause of Internet path changes** [76, 77, 78, 79, 80]. CISE-funded collaborators at UCSD and U. MD (see LOCs) have committed to use the FANTAIL system to extend their research into Internet outages and hijacks (CISE collaborations CNS-1228994, CNS-1423659, and OAC-1848641).

Despite its strategic importance, the research community lacks rigorous methods to analyze how a **country's or region's infrastructure may be vulnerable to targeted attacks by virtue of their interconnection topology**. A CISE-funded CAIDA collaboration with the University of Wisconsin is taking the first step toward identifying important components of the Internet topology of a country/region – Autonomous Systems (ASes), Internet Exchange Points (IXPs), PoPs, colocation facilities, and physical cable systems that represent "key terrain" in cyberspace. This endeavor requires a multi-layer mapping effort to discover the key components, relationships between them, and their geographic properties. A related collaboration with Internet Initiative Japan (IIJ) i developing methods and metrics to capture the relative importance of specific networks and links for connectivity to a given country and its address space. IIJ has examined this question through the lens of BGP data only [81]. our collaboration (see IIJ LOC) will adapt these methods to traceroute data, which requires searching for specific IXPs, prefixes, geographic information, and across specific times, to refine our picture of strengths and weaknesses of the Internet as critical infrastructure (CISE CNS-1705024). Another CISE-funded collaborator Rob Beverly will use FANTAIL to explore single points of failure in Internet regions (see LOC NPS).

Incidentally, another promising use of the FANTAIL system – given the use of country annotations for VPs and targets – is to facilitate refinement of methods for geolocation of Internet

infrastructure, an ongoing challenge in the research community [82, 83, 84].

References

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

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

[3] Emile Aben, "Measuring More Internet with RIPE Atlas," January 2016. https://labs.ripe.net/Members/emileaben/measuring-more-internet-with-ripe-atlas.

[4] S. Dent, "Google is Testing Internet Speeds Straight from Search," 2015. https://www.engadget.com/2016/06/29/google-is-testing-internet-speeds-straight-from-search/.

[5] M. Luckie, A. Dhamdhere, K. Claffy, and D. Murrell, "Measured Impact of Crooked Traceroute," *ACM Computer Communications Review*, January 2011.

[6] 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.

[7] M. Gunes and K. Sarac, "Resolving IP aliases in building traceroute-based Internet maps," Technical Report UTDCS-62-06, University of Texas at Dallas, 2006.

[8] "An open-source IP alias resolution tool implemented in java and plpgsql ."

[9] 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.

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

[11] Y. Vanaubel, J.-J. Pansiot, P. Merindol, and B. Donnet, "Network Fingerprinting: TTL-based Router Signatures," in *ACM SIGCOMM Internet measurement Conference (IMC)*, Oct 2013.

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

[13] 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.

[14] M. Luckie, A. Dhamdhere, B. Huffaker, D. Clark, and k. claffy, "bdrmap: Inference of Borders Between IP Networks," in *Internet Measurement Conference (IMC)*, pp. 381–396, Nov 2016.

[15] A. Marder, M. Luckie, A. Dhamdhere, B. Huffaker, J. Smith, and k. claffy, "Pushing the Boundaries with bdrmapIT: Mapping Router Ownership at Internet Scale," in *Internet Measurement Conference (IMC)*, pp. 56–69, Nov 2018.

[16] I. Elastic, "Elasticsearch." https://www.elastic.co/products/elasticsearch.

[17] I. Apache, "Spark analytics engine." https://spark.apache.org/.

[18] "San Diego Supercomputer Center HPC Systems." https://www.sdsc.edu/services/hpc/hpc_systems.html.

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

[20] P. S. Foundation, "Python." https://www.python.org.

[21] Facebook, "Rocksdb." https://rocksdb.org/.

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

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

[24] 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.

[25] 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.
[26] kc claffy, "CAIDA participation in IPv6 day," June 2011. http://blog.caida.org/best_available_data/2011/06/05/caida-participation-in-ipv6-day/.
[27] 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.
[28] Center for Applied Internet Data Analysis (CAIDA), "The IPv4 Prefix-Probing Traceroute Dataset." http://www.caida.org/data/active/ipv4_prefix_probing_dataset.xml.
[29] Center for Applied Internet Data Analysis (CAIDA), "AS links." http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml.
[30] Center for Applied Internet Data Analysis (CAIDA), "Prefix to AS mappings." http://www.caida.org/data/routing/routeviews-prefix2as.xml.
[31] Center for Applied Internet Data Analysis (CAIDA), "AS Taxonomy." http://www.caida.org/data/active/as_taxonomy/.
[32] CAIDA, "AS links annotated with AS relationships dataset." http://www.caida.org/data/active/as-relationships/index.xml.
[33] CAIDA's Macroscopic Internet Topology Data Kit (ITDK). http://www.caida.org/data/active/internet-topology-data-kit/.
[34] CAIDA Routed /24 DNS Names Dataset. http://www.caida.org/data/active/ipv4_dnsnames_dataset.xml.
[35] V. Jacobson, "traceroute." ftp://ftp.ee.lbl.gov/traceroute.tar.gz.
[36] M. Luckie, "Scamper: a scalable and extensible packet prober for active measurement of the Internet," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2010.
[37] E. C. Rosen, Y. Rekhter, D. Tappan, D. Farinacci, G. Fedorkow, T. Li, and A. Conta, "MPLS Label Stack Encoding." Internet Draft: draft-ietf-mpls-label-encaps-08.txt (expires March 2000), July 2000.
[38] Yves Vanaubel and Jean-Romain Luttringer and Pascal Mrindol and Jean-Jacques Pansiot and Benoit Donnet, "TNT, Watch me Explode: A Light in the Dark for Revealing MPLS Tunnels."
[39] CAIDA Internet eXchange Points (IXPs) Dataset. http://www.caida.org/data/ixps/.
[40] CAIDA Border Mapping Dataset. http://www.caida.org/data/active/bdrmap_dataset.xml.
[41] I. Livadariu, A. Elmokashfi, A. Dhamdhere, and kc claffy, "A first look at IPv4 transfer markets," in *CoNEXT*, 2013.
[42] "CAIDA Internet Data." http://www.caida.org/data/.
[43] 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.
[44] 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.
[45] kc claffy, "The 3rd Workshop on Active Internet Measurements (AIMS-3) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 41, July 2011.
[46] kc claffy, "The 4th Workshop on Active Internet Measurements (AIMS-4) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 42, Jul 2012.

[47] kc claffy, "The 5th Workshop on Active Internet Measurements (AIMS-5) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 43, Jul 2013.
[48] kc claffy, "The 6th Workshop on Active Internet Measurements (AIMS-6) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 44, Oct 2014.
[49] kc claffy, "The 7th Workshop on Active Internet Measurements (AIMS-7) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 46, Jan 2016.
[50] kc claffy, "The 8th Workshop on Active Internet Measurements (AIMS-8) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 46, Oct 2016.
[51] kc claffy, "The 9th Workshop on Active Internet Measurements (AIMS-9) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 47, Oct 2017.
[52] k. claffy, "Workshop on Internet Economics (WIE2009) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 40, Apr 2010.
[53] k. claffy, "Workshop on Internet Economics (WIE2011) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 42, pp. 110–114, Apr 2012.
[54] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2012) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, 2013.
[55] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2013) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 44, pp. 116–119, Jul 2014.
[56] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2014) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 45, pp. 43–48, Jul 2015.
[57] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2015) Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jul 2016.
[58] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2016) Final Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jul 2017.
[59] k. claffy, G. Huston, and D. Clark, "Workshop on Internet Economics (WIE2017) Final Report," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 48, p. 4, Jul 2018.
[60] k. claffy and D. Clark, "Workshop on Internet Economics (WIE2018) Final Report," *ACM SIGCOMM Computer Communication Review (CCR)*, Jan 2019.
[61] k claffy, "CAIDA Annual Report 2017 - Data," 2017. `http://www.caida.org/home/about/annualreports/2017/#data`.
[62] Center for Applied Internet Data Analysis (CAIDA), "Papers published (by non-caida authors using caida datasets." http://www.caida.org/data/publications/.
[63] Center for Applied Internet Data Analysis (CAIDA), "Papers Published (by CAIDA Authors." http://www.caida.org/publications/papers/.
[64] Jean-Franois Grailet, Benoit Donnet, "Towards a Renewed Alias Resolution withSpace Search Reduction and IP Fingerprinting," 2017. `https://orbi.uliege.be/bitstream/2268/209979/1/paper.pdf`.
[65] I. Cunha, P. Marchetta, M. Calder, Y. Chiu, B. Schlinker, B. Machado, A. Pescap, V. Giotsas, H. Madhyastha, and E. Katz-Bassett, "Sibyl: A practical internet route oracle," in *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI '16)*, 2016.
[66] 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.
[67] D. Clark, S. Bauer, k. claffy, A. Dhamdhere, B. Huffaker, W. Lehr, and M. Luckie, "Measurement and Analysis of Internet Interconnection and Congestion," in *Telecommunications Policy Research Conference (TPRC)*, Sep 2014.

[68] A. Dhamdhere, D. Clark, A. Gamero-Garrido, M. Luckie, R. Mok, G. Akiwate, K. Gogia, V. Bajpai, A. Snoeren, and k. claffy, "Inferring Persistent Interdomain Congestion," in *ACM SIGCOMM*, Aug 2018.

[69] E. Katz-Bassett, H. V. Madhyastha, J. P. John, A. Krishnamurthy, and T. Anderson, "Studying black holes in the Internet with Hubble," in *USENIX Symposium on Networked Systems Design & Implementation (NSDI)*, 2008.

[70] A. Schulman and N. Spring, "Pingin' in the rain," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2011.

[71] A. Dainotti, C. Squarcella, E. Aben, K. Claffy, M. Chiesa, M. Russo, and A. Pescap?, "Analysis of Country-wide Internet Outages Caused by Censorship," in *Internet Measurement Conference (IMC)*, (Berlin, Germany), pp. 1–18, ACM, Nov 2011.

[72] A. Dainotti, R. Amman, E. Aben, and K. Claffy, "Extracting benefit from harm: using malware pollution to analyze the impact of political and geophysical events on the Internet," *SIGCOMM Comput. Commun. Rev.*, vol. 42, January 2013.

[73] L. Quan, J. Heidemann, and Y. Pradkin, "Towards Active Measurements of Edge Network Outages," in *Passive and Active Measurement*, 2013.

[74] L. Quan, J. Heidemann, and Y. Pradkin, "Trinocular: Understanding Internet Reliability Through Adaptive Probing," in *ACM SIGCOMM*, August 2013.

[75] X. Fan and J. Heidemann, "Selecting representative IP addresses for internet topology studies," in *ACM SIGCOMM Internet measurement Conference (IMC)*, 2010.

[76] I. Cunha, R. Teixeira, N. Feamster, and C. Diot, "Measurement methods for fast and accurate blackhole identification with binary tomography," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2009.

[77] I. Cunha, R. Teixeira, and C. Diot, "Measuring and Characterizing End-to-End Route Dynamics in the Presence of Load Balancing," in *Passive and Active Measurement Conference*, April 2011.

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

[79] 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.

[80] 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.

[81] R. Fontugne, A. Shah, and E. Aben, "As hegemony: A robust metric for as centrality," in *Proceedings of the SIGCOMM Posters and Demos*, pp. 48–50, ACM, 2017.

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

[83] 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.

[84] 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/`.