

caida

Recent and Future Work

cooperative association for internet data analysis

Bradley Huffaker <bradley@caida.org>

12 June 2006



caida Update

recent

- public release of AS Relationship data (weekly)
- public release of DatCat
- WIT workshop
- DNS anycast analysis

future/proposed work

- Next Generation Measurement Infrastructure
- “A Day in the Life of the Internet” experiment
- economics and policy research
- Future Internet Routing Architecture
- automation of DNS open resolver measurement
- upgrade backbone passive monitors to OCI92



caida Update

recent

- public release of AS Relationship data (weekly)
- public release of DatCat
- WIT workshop
- DNS anycast analysis

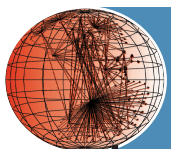
future/proposed work

- Next Generation Measurement Infrastructure
- “A Day in the Life of the Internet” experiment
- economics and policy research
- Future Internet Routing Architecture
- automation of DNS open resolver measurement
- upgrade backbone passive monitors to OCI92



Recent

**caida's work since last WIDE
workshop (18 mar 2006)**



caida

DatCat



<http://imdc.datcat.org>

 **DatCat** is an **NSF**
funded Internet measurement
data catalog.



Target Problems



<http://imdc.datcat.org>

- data everywhere and lots of it
 - caida alone has over 50 terabytes of data
 - pcap (tcpdump) packet traces
 - skitter topology traces
 - routing tables
 - etc
- data is hard to find
 - no central indexing
 - many one-off data collections

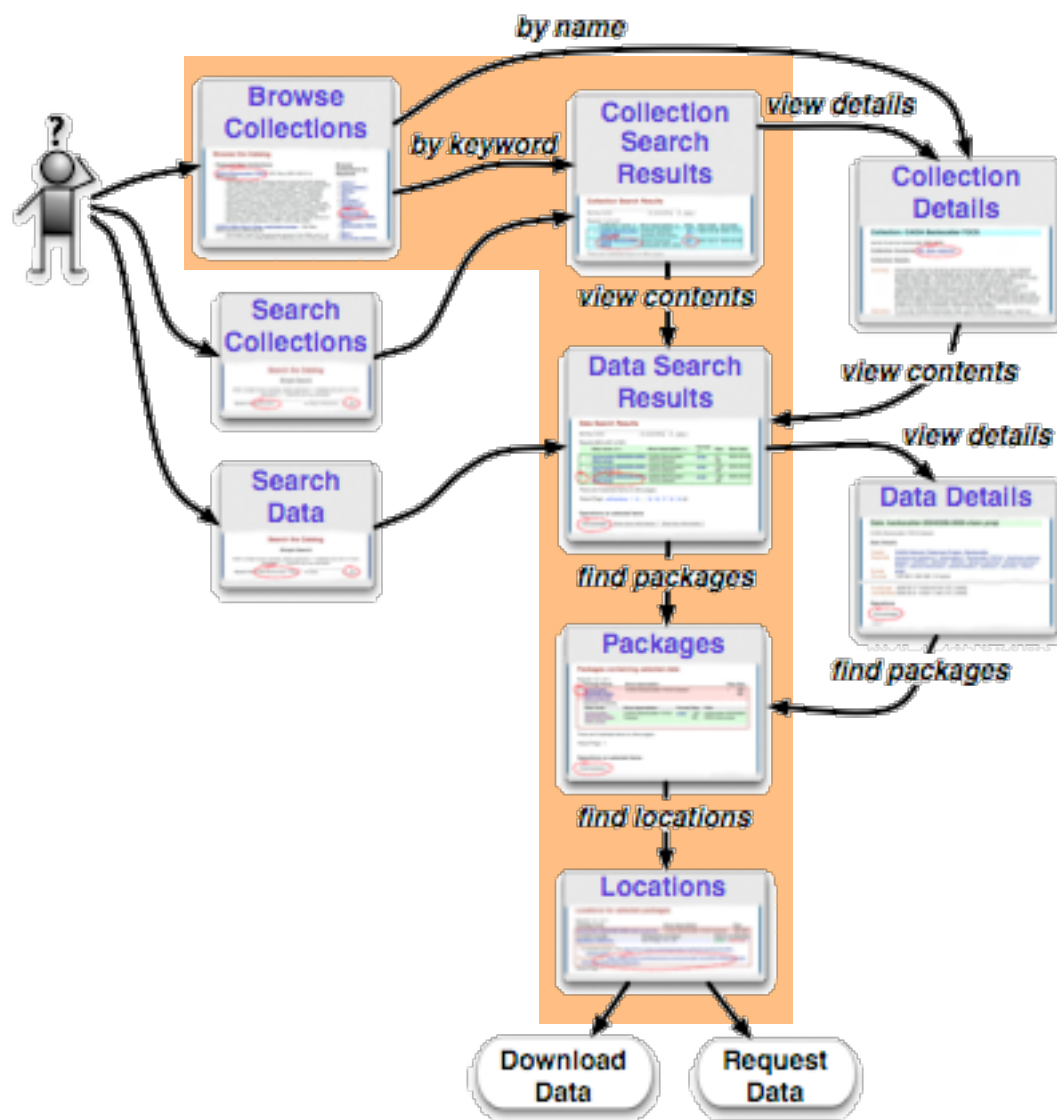
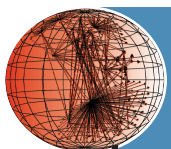
Make the following processes easy for users.

- finding data sets of interest
 - Many researchers lack access and/or expertise to collect data needed for their research.
- adding new data sets to the catalog
 - Contributors (who are generally underfunded and providing data out of dedication to the general good) want to minimize time lost to their own research.
- annotating data sets in the catalog
 - Provides a flexible way for contributors and users to mark up interesting facts about data sets. Such as the number of packets or that a given file is corrupted.



current implementation

- user accounts
- data sets
 - only CAIDA data so far
 - 57,088 files indexed
 - 4.56 TB of data indexed
 - 11 separate collections
- browse simple/advanced search
- help/tutorial
- API for bulk contributions





Browse Catalog



<http://imdc.datcat.org>

Browse the Catalog

Featured Data Collections

[CAIDA Backscatter-TOCS](#) - 231 files, 2001-02-01 to 2004-03-06
Information useful for studying denial-of-service (DoS) attacks. This dataset consists of 3 billion IPv4 packets sent by DoS attack victims in response to spoofed attack traffic. This backscatter from victims was collected by the UCSD Network Telescope. Possible uses include modeling DoS attacks, understanding victim populations, and using real packet traces to validate algorithms for detecting or classifying malicious traffic. This last use is particularly valuable because it is extremely challenging to artificially generate the kind of real-world noise present on the Internet. This dataset includes just the subset of CAIDA's backscatter data used for the paper "Inferring Internet Denial-of-Service Activity" published in ACM TOCS, May 2006.

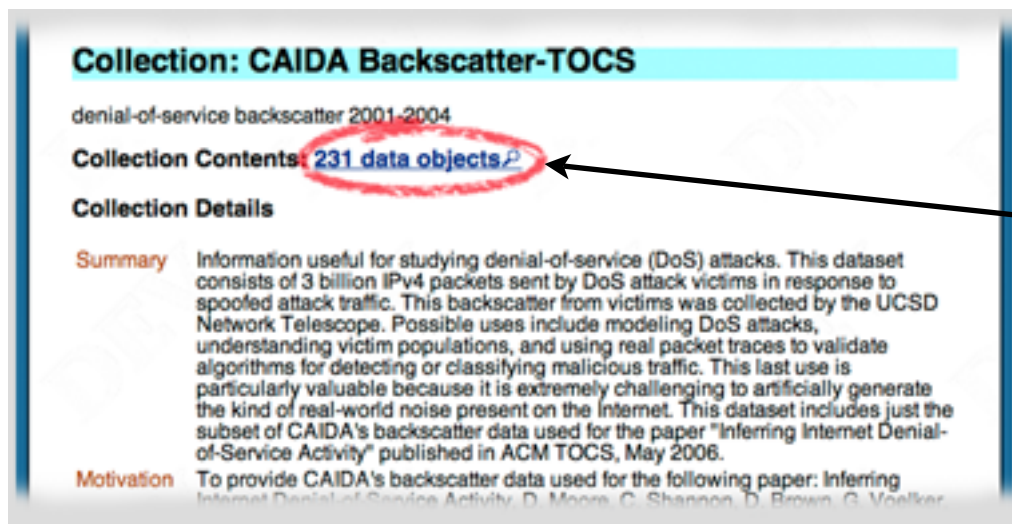
[CAIDA Witty Worm Data, restricted access](#) - 132 files, 2004-03-20 to 2004-03-25
Information useful for studying the spread of the Witty worm, as observed by the UCSD Network Telescope over a 5-day period

Browse Collections by Keyword

- [active](#)
- [anonymized](#)
- [ARTS](#)
- [AS](#)
- [AS links](#)
- [background radiation](#)
- [backscatter](#)
- [Backscatter-2004-2005](#)
- [Backscatter-TOCS](#)
- [BGP](#)
- [blackhole address](#)

select collection of interest

- Browse page is a list of data file collections.
 - Collections are a high level group of related files.
 - A file may belong to more than one collection.
- User clicks on collection of interest to view the [collection details](#).



Collection: CAIDA Backscatter-TOCS

denial-of-service backscatter 2001-2004

Collection Contents: [231 data objects](#)

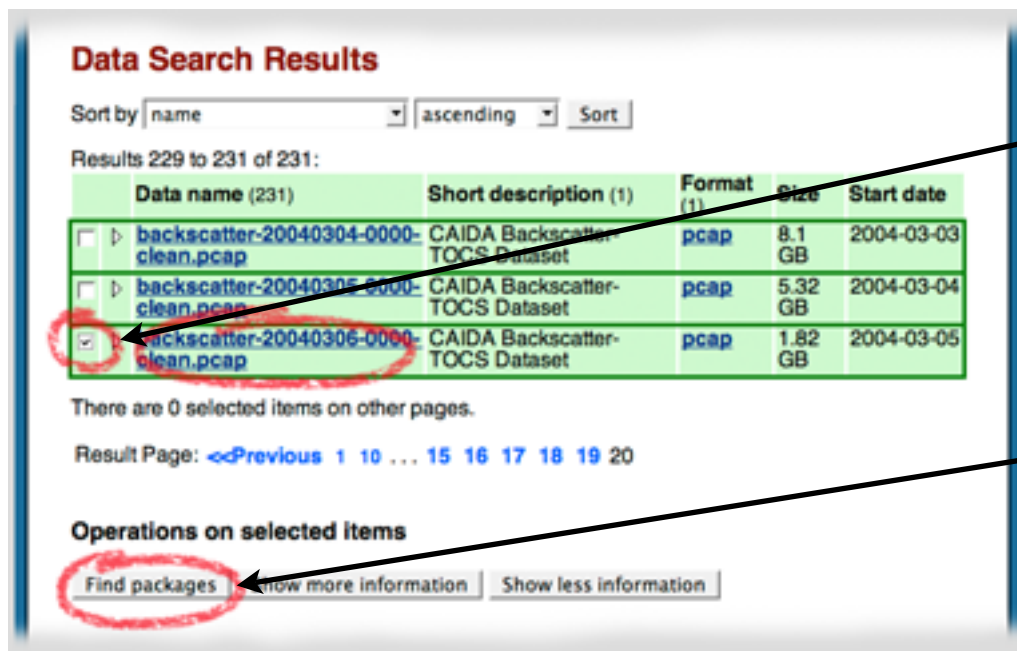
Collection Details

Summary Information useful for studying denial-of-service (DoS) attacks. This dataset consists of 3 billion IPv4 packets sent by DoS attack victims in response to spoofed attack traffic. This backscatter from victims was collected by the UCSD Network Telescope. Possible uses include modeling DoS attacks, understanding victim populations, and using real packet traces to validate algorithms for detecting or classifying malicious traffic. This last use is particularly valuable because it is extremely challenging to artificially generate the kind of real-world noise present on the Internet. This dataset includes just the subset of CAIDA's backscatter data used for the paper "Inferring Internet Denial-of-Service Activity" published in ACM TOCS, May 2006.

Motivation To provide CAIDA's backscatter data used for the following paper: Inferring Internet Denial-of-Service Activity, D. Moore, C. Shannon, D. Brown, G. Voelker

go to a list of collection's data objects

- Collection details displays specifics of the collection.
 - description of collection and its contents
 - motivation behind the collection
- If the collection matches the researchers' interest they can then [list the data objects](#).



Data Search Results

Sort by

Results 229 to 231 of 231:

Data name (231)	Short description (1)	Format (1)	Size	Start date
<input type="checkbox"/> backscatter-20040304-0000-clean.pcap	CAIDA Backscatter-TOCS Dataset	pcap	8.1 GB	2004-03-03
<input type="checkbox"/> backscatter-20040305-0000-clean.pcap	CAIDA Backscatter-TOCS Dataset	pcap	5.32 GB	2004-03-04
<input checked="" type="checkbox"/> backscatter-20040306-0000-clean.pcap	CAIDA Backscatter-TOCS Dataset	pcap	1.82 GB	2004-03-05

There are 0 selected items on other pages.

Result Page: <<Previous 1 10 ... 15 16 17 18 19 20

Operations on selected items

select desired data

find the data's packages.

- A listing of the data files contained within the selected collection. Also shown: general statistics about the files such as format and size.
- After the user has selected the set of files they are interested in, they then **find packages** which contain them.

Packages containing selected data

Results 1 to 1 of 1:

Package Name	Short description	Files	Size
<input checked="" type="checkbox"/> backscatter-20040306-0000-clean.pcap.izo	CAIDA Backscatter-TOCS Dataset	1	566 MB

Selected contents:

Data name	Short description	Format	Size	Path
backscatter-20040306-0000-clean.pcap	CAIDA Backscatter-TOCS Dataset	pcap	1.82 GB	backscatter-20040306-0000-clean.pcap

There are 0 selected items on other pages.

Result Page: 1

Operations on selected items

select desired packages

find download locations

- Packages are the downloadable groupings of one or more data files.
- Once the user has selected a set of packages, he then **finds download locations**.

Locations for selected packages

Results 1 to 1 of 1:

Package name	Short description	Size
backscatter-20040306-0000-clean.pcap.i zo	CAIDA Backscatter-TOCS Dataset	566 MB

Location handle	Geographic location	Status	Availability
/location/1-3NCP-G	San Diego, CA, US	active	restricted

Download procedure:
to request access, visit: http://www.caida.org/analysis/security/telescope/backscatter_request.xml

Download URL: <https://data.caida.org/datasets/security/backscatter-tocs/2004-03/backscatter-20040306-0000-clean.pcap.i zo>

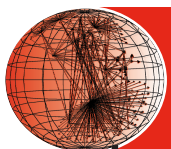
Result Page: 1

select a location for the package

- Locations are the place or process by which the package can be obtained.
 - Provides a simple URL or instructions which must be followed to get the data.
- Get your data! 😊



- small number of invited third party contributors
- public contributors
- tools
- studies
 - collections specialized for papers / experiments / etc



caida

WIT



<http://www.caida.org/workshop/isma/0605>



CAIDA
cooperative association
for internet data analysis

WIT

workshop on internet topology | UCSD | SDSC | San Diego, CA | May 2006

**Workshop on the Internet Topology (WIT)
brought together researchers from a
diverse group of communities.**



Overview

<http://www.caida.org/workshop/isma/0605>

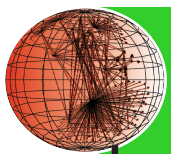
- major disagreement between physics and computer science communities (over 5 years)
 - physicists felt there was too much details
 - computer scientists felt there were too few
- brought together leading researchers from both communities.
- in the end both sides agree that this disagreement was counter productive
- agreed to collaborate in the future
- final report will be delivered summer 2006



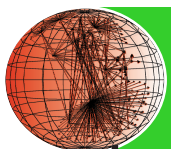
DNS anycast analysis

DNS anycast analysis

**Using tcpdump from three roots
we examined the geographic and topological
clustering of DNS clients.**



- dates
 - January 10th-11th 2006
- DNS sources
 - c-root (Cogent) 7 out of 7 instances
 - f-root (ISC) 61 out of 71 instances
 - k-root (RIPE) 24 out of 31 instances
- geographic
 - Netacuity database used for geographic mapping
- topological
 - Routeviews used for ASs and prefixes (January 10th)

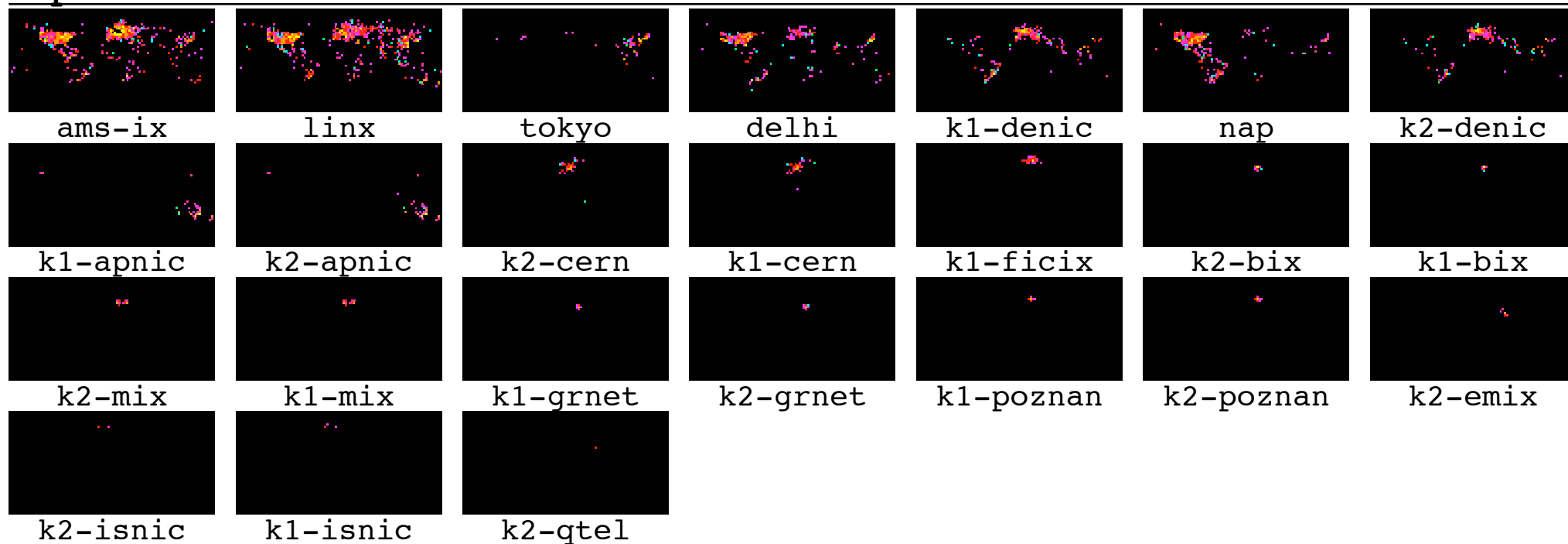


Client Locations

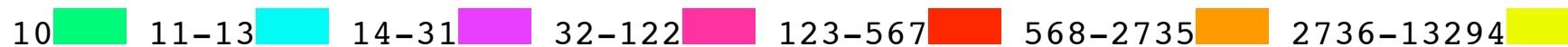
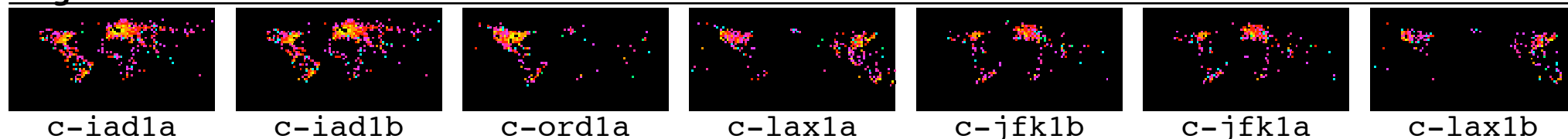
caida

DNS anycast analysis

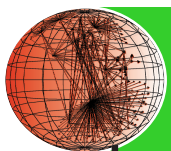
ripe



cogent



Each map is a single instance. The color represents the number of clients at a geographic location.

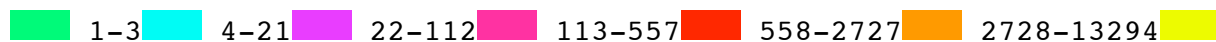
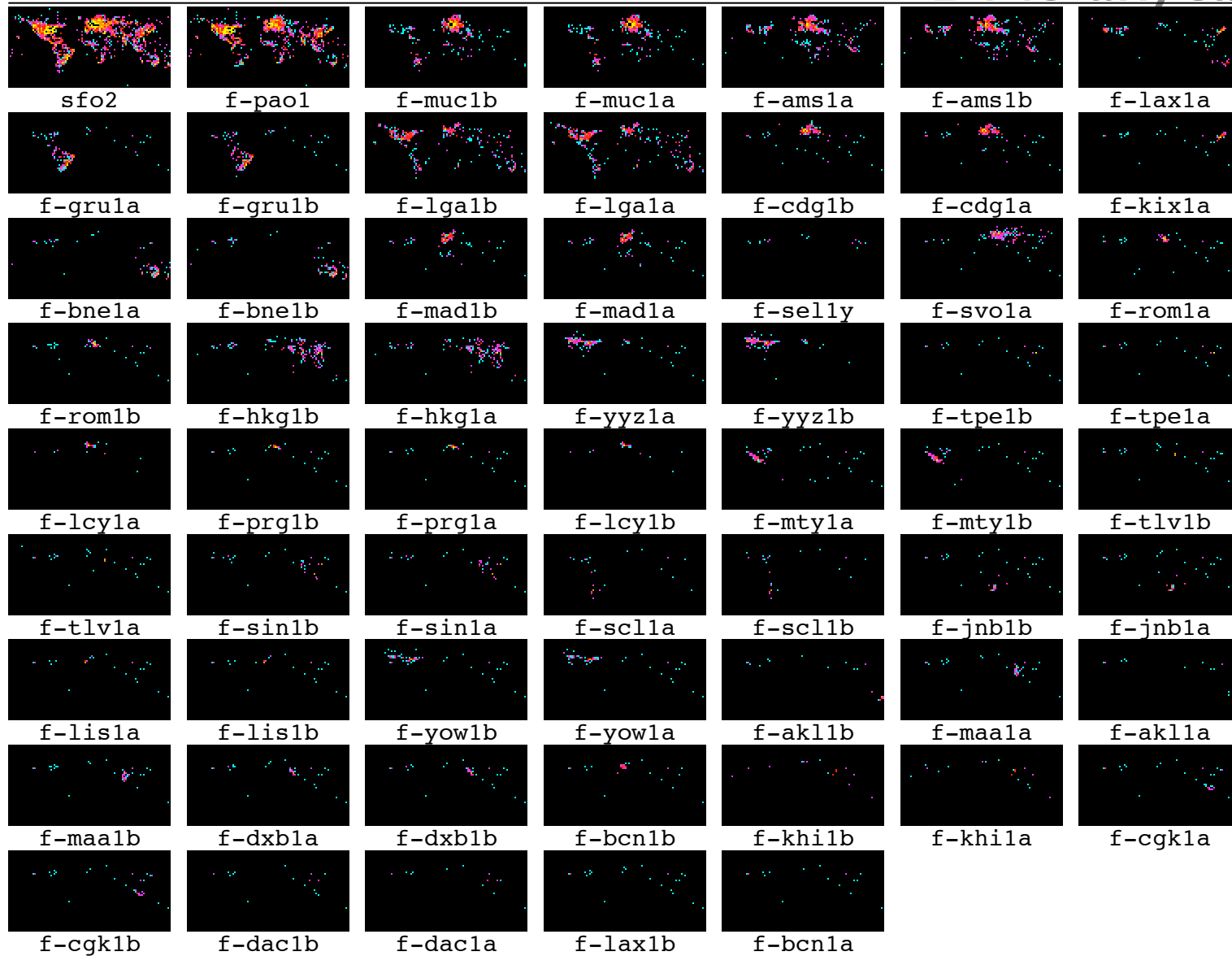


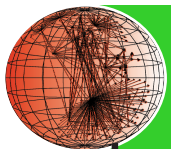
caida

Client Locations

DNS anycast analysis

isc



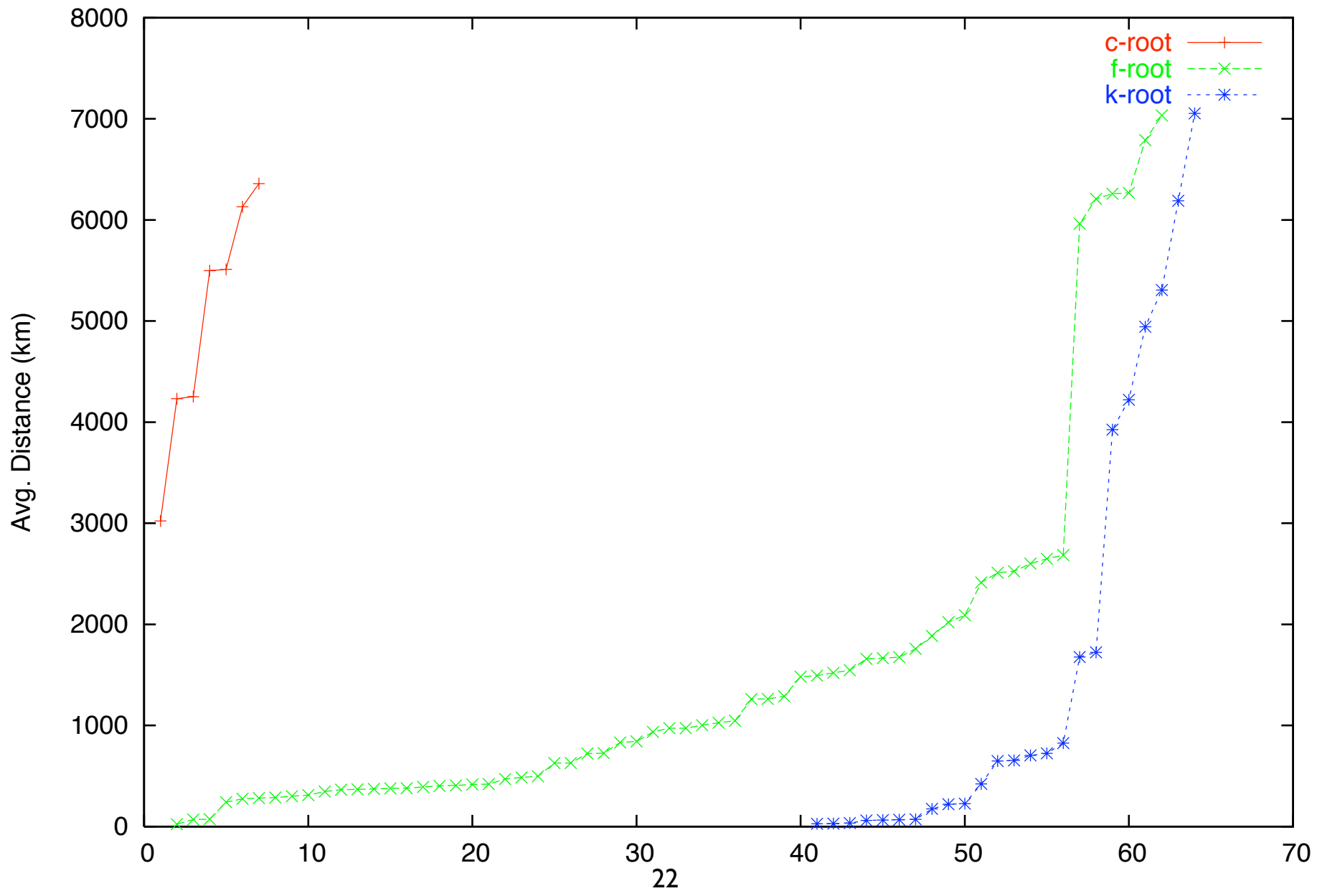


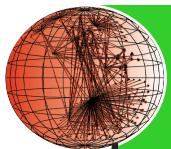
calda

Geographic Client Clustering

DNS anycast analysis

Average Geographic Distance between Hosts



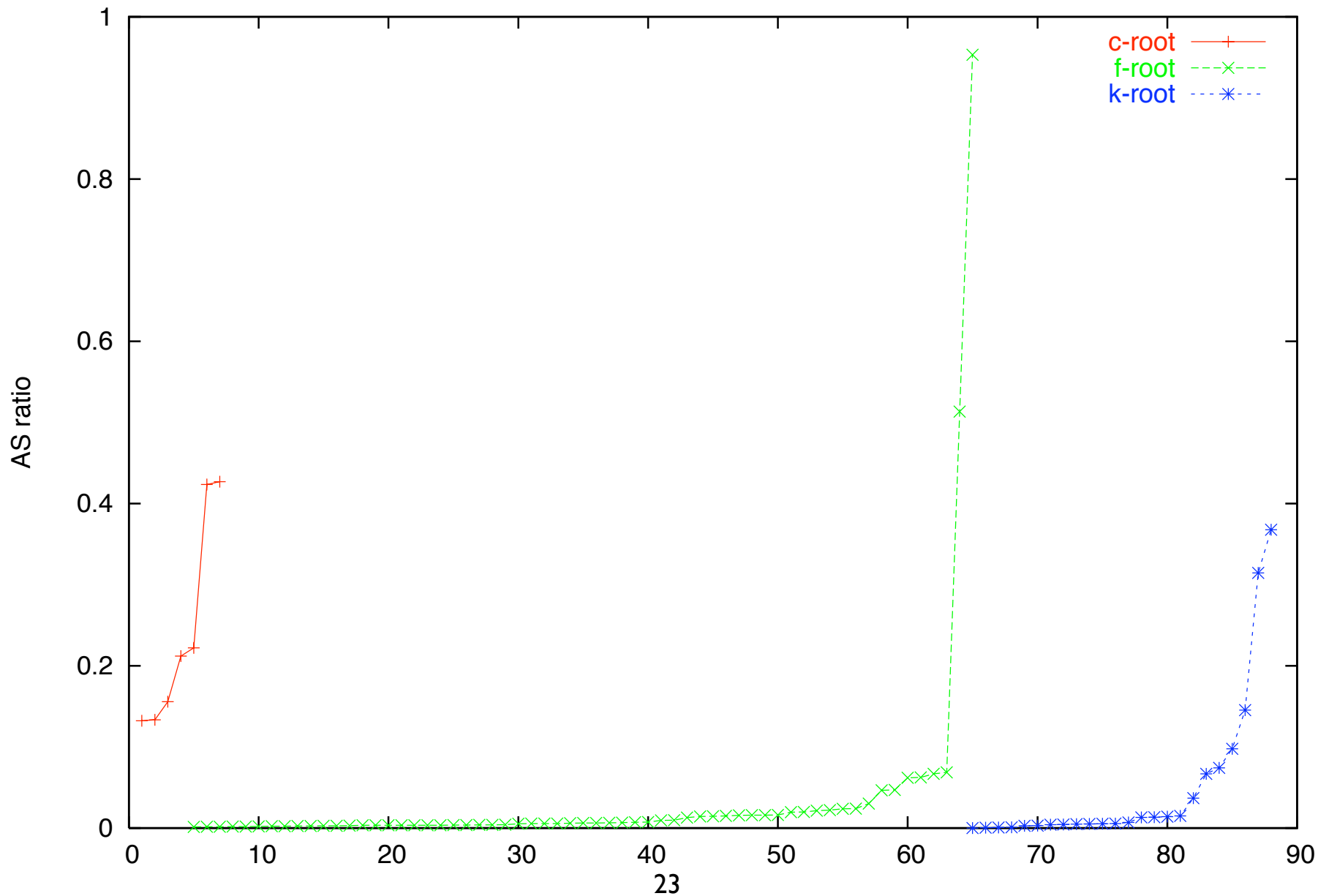


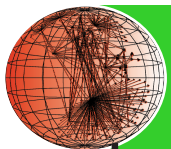
caida

Server's ASES / All ASES (observed)

DNS anycast analysis

ratio of ASES seen by each server to all ASES seen



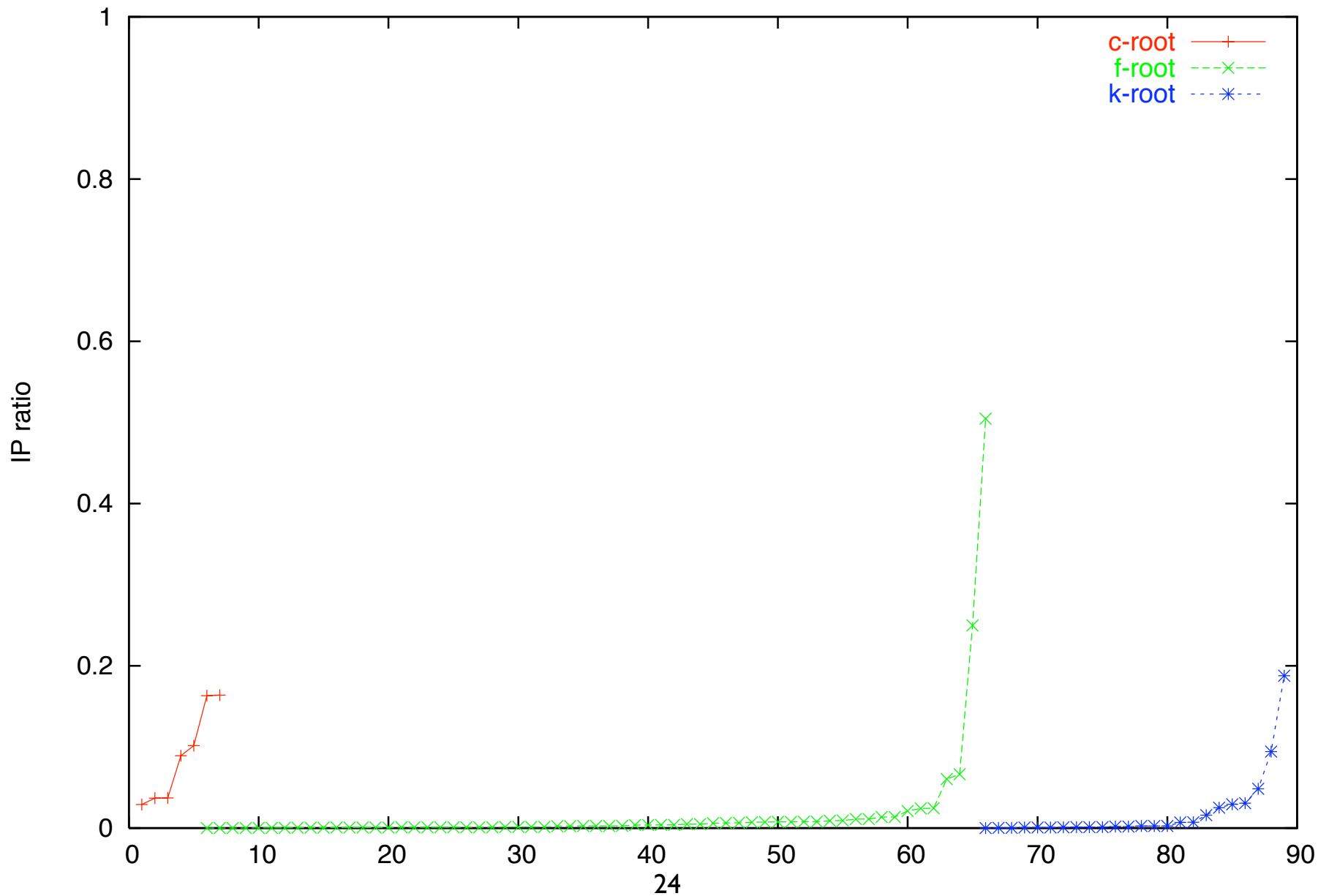


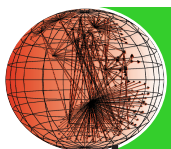
caida

Server's Clients / All Clients (observed)

DNS anycast analysis

ratio of clients seen per server to all clients seen





Summary

DNS anycast analysis

AS coverage

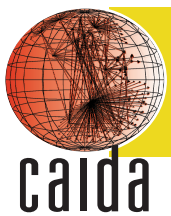
geographic coverage

c-root (Cogent)	2 around 45% 2 around 22% 3 around 16%	2 globals 5 semi-global
k-root (RIPE)	2 around 39% 5 between 5-10% 19 between 0-5%	2 global 4 semi-global 18 regionals
f-root (ISC)	1 around 95% 1 around 50% 6 between 5-10% 63 between 0-5%	2 global 9 semi-global 50 regionals



Intended Work

**caida's work from
June 12 into the future**



Measurement Infrastructure

next generation measurement infrastructure

**Next Generation Measurement Infrastructure
(NGMI) will be
scalable, flexible, and shareable.**



Goals

next generation measurement infrastructure

- **greater scalability**
 - system management, number of monitors, and measurement efficiency
- **greater flexibility**
 - deployment, type of measurements, and mixing unrelated measurements
- **decreased researcher's burden**
 - provide common, simple, and powerful APIs
 - remove/reduce OS/hardware required knowledge
 - handle inter-monitor communication



Comparison: Scriptroute

next generation measurement infrastructure

- allows “untrusted” third parties
 - our system - only screened/trusted third parties
- severely limits the type of packets
 - our system - will allow more types of packets
- limits the amounts of resources
 - our system - single process may have all available resources
- provides no central management
 - our system - both system and experiment communication channels



Comparison: DIMES

next generation measurement infrastructure

- large deployment on third party end users machines (uses idle resources on desktops)
 - our system - only dedicated systems
- severely limited resources per host (although made up for in part by numbers)
 - our system - single process may use all available resources
- possible untrustworthy participants
 - our system - only screened/trusted participants



“A Day in the Life...”

unfunded

“A Day in the Life of the Internet”

- simultaneous capture of variety of data:
 - workload, topology, routing, performance, DNS
anonymized appropriately according to local laws
 - from strategic locations around the globe
- will require interdisciplinary approaches
 - logistics, economics, and legal challenges
- will require cooperation/partners among
 - governments, ISPs, and educational networks



toward an empirical network macro-economics

- define internet public interest
- identify empirical data resources to evaluate defined terms
- empirically evaluate explanations of macroeconomic behavioral patterns and anomalies
- identify economic risks of policy strategies



conclusion

recent

- DataCat

<http://imdc.datcat.org>

- WIT workshop

<http://www.caida.org/workshop/isma/0605>

final report by July 2006

- DNS anycast analysis

future work

- Next Generation Topology Infrastructure

- Day in the Life of the Internet