

Data Collection/Provision at CAIDA

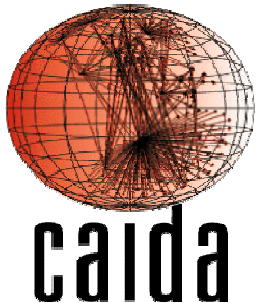
Colleen Shannon (CAIDA)

cshannon @ caida.org

dmoore @ caida.org

kc @ caida.org

www.caida.org



Outline

- Partners / Data Sources
- Data Sets
 - Previously collected
 - On-demand collections
- Collection Process and Architecture
 - Collectors
 - Data storage
 - Data registration
 - Data distribution
- Internet Measurement Data Catalog



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Partners / Data Sources

- PAIX (OC48 peering links)
- Equinix (OC48 and GigE peering links)
 - 100 potential providers
- UCSD
- Network Telescope
- 22 Active Monitors spread across 5 continents



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Data Collection – Current Data Sets

- Anonymized OC48 traces from the past year
 - Data researchers can get started on now
 - Historical context for trend analysis and targeted tool development
- Denial-of-Service backscatter traces
 - 22 traces of denial-of-service activity from January 2001
 - February 2004
- Witty Internet worm data set
- Network Topology data
 - AS adjacencies
 - Hop-by-hop topology traces (skitter)



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Data Collection – On-Demand Collection

- Some preset traces
- Peering link traces (OC48, GigE)
 - Anonymized? Summarized? Which site?
- Network telescope data
 - Near-continuous collection
 - researchers request specific intervals
- Network topology data



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Collection Process & Architecture

- CAIDA is both a data provider and a data hosting site
 - Collection and hosting of data we own
 - Collection and hosting of data from our industry partners
- Request and Review
- Collectors
- Data Storage
- Data registration
- Data distribution



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Data Storage

- All data can't be instantly available
 - Network telescope collects 35GB of data every day...
- All data can't be available simultaneously
 - Researchers download data in planned time windows
 - Summarized data more ubiquitously available
- Site repository security is a high priority



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Data Registration

- How do researchers know if a dataset is useful?
 - Have to jump through hoops to access the data
 - Need information on collection location, format, size, data features, etc.
- Public data sets registered with Internet Measurement Data Catalog
 - Access information
 - Ownership information
 - Annotation



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Data Distribution

- Secure distribution of large data sets is a nontrivial problem
 - Bandwidth constraints
 - Tool/access constraints
 - Physical device exchange (Fedex can be very high bandwidth...)
- Funding model
 - Researchers provide hardware
 - Shipping costs



Internet Measurement Data Catalog

- Goals
 - Let researchers know what data is available
 - Access is difficult but finding out what's out there and where it is is still 90% of the battle
 - Specific characteristics of the data often determine its utility
 - Let researchers know how and to whom data is available
 - Who do you ask?
 - When should you bother asking?
 - What is the AUP?
 - Goal is **not** to host/serve data



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Internet Measurement Data Catalog

- Architecture
 - Database
 - Adequately describing datasets is hard
 - Consumers want all possible information, providers want to provide as little information as possible
 - Web Interface
 - Sophisticated searching
 - Navigable data selection/listing
 - API
 - Providers need to be able to submit data automatically



IMDC – Problems to Solve

- What is the data?
 - What format is it?
 - Where is it from?
 - What's wrong with it?
 - What useful features does it have?
 - Who has it?
 - How do I get it?
 - How do I read it/process it?
 - Who do I talk to when I have problems?



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Overall Goals

- Provide as much raw, unfiltered data to researchers as possible to facilitate good research and development of useful tools
- Protect network and system users
 - Privacy
 - Security
- Accomplish this as smoothly, quickly, transparently, and fairly as possible



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE

Acknowledgements

- CAIDA folks
- Cisco Systems – initial investment in data collection/provision infrastructure
- PAIX
- Verio
- Equinix
- NSF



COOPERATIVE ASSOCIATION FOR INTERNET DATA ANALYSIS

University California, San Diego – Department of Computer Science



UCSD-CSE