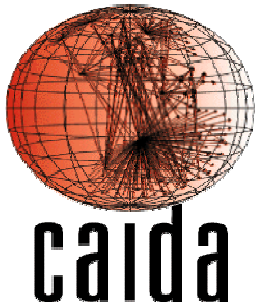


Data Catalog ISMA: CAIDA Data Collection

Colleen Shannon (CAIDA)

cshannon @ caida.org

www.caida.org



Outline

- Current CAIDA project areas
- CAIDA data collection
 - Passive data collection (me)
 - Active/topology data collection (brad)
 - DNS and routeviews and other data (k)



Current Project Areas

- Routing topology and behavior
- Passive monitoring and workload characterization
- Internet Measurement Data Catalog
- Bandwidth estimation
- Flow collection and efficient aggregation
- Security: DoS and Internet worms
- DNS performance and anomalies
- Visualization



Current Project Areas

- Routing topology and behavior
 - Skitter, scamper; monitors around the world
- Internet Measurement Data Catalog
- Trace Collection and Storage
 - Maintaining remote monitors
 - Transferring files back to SDSC
 - Sanitizing data
 - Managing data access
- Security: DoS and Internet worms



Tools

- CoralReef & NeTraMet
- Walrus & Otter, libsea, PlotPaths
- NetGeo
- Skitter
- Graph::Chart, GeoPlot,
- ASFinder
- Beluga, GTrace
- dnstat, dnstop
- Collaborations:
 - RRDTool, AutoFocus, PathRate/PathLoad



Passive Data Collection

- Data sources
- Data processing
 - Aggregation
 - Context and summary data generation
- Data management/storage
- Providing data access



Passive Data Sources

- Peering links for backbone service providers
 - OC12 and OC48
 - Provider-specific collection agreements
- UCSD campus access links
 - Some collection freedom due to campus support for research
 - Many additional restrictions imposed by the presence of students and student data
- DNS root servers
- UCSD Network Telescope
 - Traffic initiated by external sources; few restrictions on data use
 - Profound security concerns about the release of some data



Network Telescope

- Chunk of (globally) routed IP address space
 - 16 million IP addresses
- Little or no legitimate traffic (or easily filtered)
- Unexpected traffic arriving at the network telescope can imply remote network/security events
- Generally good for seeing explosions, not small events
- Depends on random/chance component in observed events



Data Processing: Summaries

- Raw data is useful to a lot of people
- However...what raw data do researchers want?
- Aggregation provides summaries that allow researchers to identify regions of interest in large datasets
- Graph summaries let researchers quickly scan large volumes of data for interesting intervals
 - e.g. Peak utilization, unusual features
 - Context and summary data generation



Data Processing: Context Data Generation

- Raw and summarized data are useful
- External data sources can be invaluable assets
 - DNS hostname lookups (time sensitive)
 - Routing tables
 - Log of related problems/events



Data Management/Storage

- Data files are large!
 - Compression algorithms have space/processing tradeoffs
 - How active is use? What cost do various compression algorithms have?
- Continuous collection requires lots of time/patience/valium
 - Network Telescope collects ~35GB of compressed data a day
 - Even “automated” processes need to be closely managed



Providing Data Access

- Data sanitation
 - Anonymization
 - Payload stripping
- Access policies
- Serving stored data
 - Secure access to large files is difficult!
- Tracking data access
- Supporting data users
 - If you build it, they will come
 - ...and bring their questions



Conclusions

- Everyone wants data
- Collecting data is hard
 - Collecting context data is harder
- Storing data is hard
- Figuring out access policies for data is hard
- Providing data is hard
- All of the above is time consuming
- Anything that reduces the time and effort it takes to provide data to the community is useful!

