

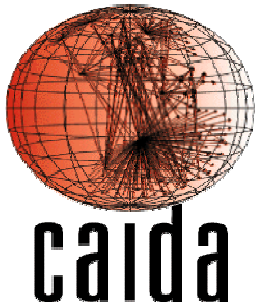
Internet Measurement Data Catalog: Motivation and Design Principles

Colleen Shannon

David Moore

cshannon, dmoore @ caida.org

www.caida.org



Problems

- There's a lot of data out there
 - Canonical: pcap (tcpdump) packet trace
 - Routing tables
 - Traceroute-type logs
 - Security data: from syslog to Network Telescope (black hole, iSink, darknet) traces
 - Names: DNS hostnames for IP addresses
 - Geographic: location mappings for IP addresses
 - Etc.
- There's a lot of weird data out there
 - How do you represent the research importance of data?
 - When the full importance isn't realized by its collector?



Motivation

- M. Allman, E. Blanton, and W. Eddy, "A Scalable System for Sharing Network Measurements." In Passive and Active Measurement (PAM) 2002.
- <http://www.icir.org/mallman/papers/simr-pam2002.ps>



Goals: General

- Main goal of Internet Measurement Data Catalog: to be an easy way for users to provide data and contributors to publish data
- Users and contributors perpetually in conflict:
 - Users want 100% complete, 100% accurate information freely available 100% of the time
 - Contributors (who are generally unfunded and providing data out of dedication to the general good) want to minimize time lost to their own research and spend as little time/money as possible providing data



Goals (2): Database Design

- Design goal 1: to require just enough information from the Contributors to make the catalog useful, while providing a framework that lets Contributors easily add additional information
- Design goal 2: to make it easy for Users to perform both simple and sophisticated searches for data
 - Good query page layout
 - Database design with explicit fields for things users will want to search for



Goals (3): User Support

- Help users share information they discover in the data
 - Supporting folks using data is a non-trivial cost for data providers
 - Original creator of dataset may not be available
- Give users the ability to correct incorrect information associated with data



Caveats

- The IMDC will not store data.
 - We have enough trouble gathering resources to store our own data. 😊
 - The technical and especially legal logistics involved in owning, storing, and serving data are untenable.
- Data entered into the IMDC is not required to be publicly available (or even available at all).
 - Documentation of datasets is the first step towards reproducibility of research results (and real science!)
 - We do hope this project encourages more folks to publish the data they use in their research.



General Design Principles

- Be ambitious: try to come up with all kinds of uses of this system
 - Not all of them will be realized, but we want to have in mind things that might be important later
- Start simple in building the system, then work up to full functionality, using feedback from how people initially use it
- Multiple access modes a necessity
 - Web, API, XML import/export

