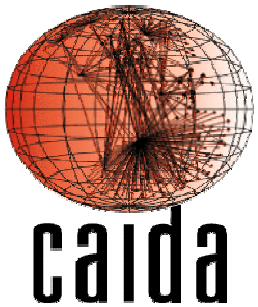
"current problems in data collection and analysis"

Colleen Shannon

cshannon @ caida.org

www.caida.org/data/



CONMI - March 30, 2005

Network Research Infrastructure

- What is “network research infrastructure”?
  - Anything from prototype router deployed in a current network to a research only network to hardware and software to monitor existing traffic
- Depends on who you are and what you want to do
- Our future as a science depends on shared infrastructure and gathered data
  - Repeatable experiments, reproducible results

Network Research

- At least three major modes of network research where infrastructure is concerned:
  - studying what is happening now (it's never what we think)
    - The problem with studying what is happening now is that it's hard! Much easier to come up with something new than to fix something broken when we don't know what is broken
  - trying out completely new things on existing infrastructure
    - what operators will let you play with critical infrastructure that is already losing money for them without potentially introducing new problems?
  - development/testing of new infrastructure
    - how do you get real traffic to test new devices?
    - how do you even simulate real traffic without knowing what it is?
    - how do you predict future trends and plan for them without knowing what is happening now?

Challenges in Data Collection for Network Research

- Access
  - Getting access to data is difficult and often depends on personal relationships
  - Privacy concerns
  - AUP concerns
  - Legal concerns
    - Student data
    - Is network monitoring legal?

Challenges in Data Collection for Network Research

- However some data exists -- so some access for successful data collection does exist

- So why isn't there more data around to work on?
  - Cost
  - Funding
  - Incentive
  - Legal Problems
  - Distribution
  - Ongoing support

Challenges in Data Collection -- Cost

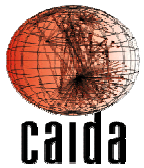
- Data collection is expensive
  - And the complexity of the task is not as intuitive as why it's expensive to build and maintain a physical object e.g. research ship, large telescope
- Storing data is expensive
  - Networks getting faster; traffic increasing
  - e.g. Network telescope generates 35-40GB of compressed data/day, which is ~10TB/year
- Dearth of hardware for specialized measurement
  - Only a few vendors
  - Expensive to develop, test, and produce
- Deploying and maintaining hardware is costly and time-consuming
  - More difficult to support remote machines than local ones
- Cost pushes researchers away from important problems

Challenges in Data Collection -- Funding

- Data collection is an engineering effort that supports research
- Lots of mechanisms for funding research, but few for supporting the necessary data
  - Funding bodies want to fund results much more than infrastructure

Challenges in Data Collection -- Incentive

- For the people who have taken the time/resources to collect data, what incentive do they have to make that available?
- High cost:
  - Months of time to prepare so it can be useful and support over time
  - Distribution can be difficult
  - Requires frank appraisal of data problems
- Little incentive
  - Professional assessment and accolades don't acknowledge and reward making data available



Challenges in Data Collection -- Legal Problems

- The question of whether it is legal to collect data on a network that carries legitimate traffic in this country is not resolved.
  - No one owns the stars (yet)
- Carriers are not protected in their ability to collect operational/research data
  - Can be subpoenaed
  - RIAA etc. would leap at that chance
- Privacy concerns are significant
- Many collections made “unofficially”
  - Distribution requires much wider acknowledgement that a collection exists

Challenges in Data Collection -- Distribution

- Giving out a terabyte of data is tough
  - Especially if security of the data is a concern
- Significant cost
  - Servers and other hardware
  - People to support hardware and handle data distribution
  - People to resolving problems and answer questions about data
- Trust
  - Most data has privacy concerns
  - How do you only hand out data to the “good guys”?

Challenges in Data Collection -- Maintenance

- Continued operation and distribution is difficult to fund
  - "that's already been done"
- Extremely difficult to do research on trends because there is little historical data
- Without community-supported collections and archives, this is not likely to change

Impacts

- The difficulty of gaining access to data is pushing solving current network problems out of scope for many projects
- We need community support so we don't continue to lose ground on our ability to monitor networks
- We need to make data sharing a requirement for publication (after waiting period)
  - Protein Data Bank success driven by this requirement
  - Pushes costs into proposals