

How Do We Build A Culture That Values Data Catalogs?

Mark Allman
ICSI / ICIR
ma11man@icir.org

CAIDA Data Catalog Workshop
June 2004

"Hit a bump and somebody screamed, you shoulda heard just what I'd seen"

Sharing Measurements

- Why share?
 - ▶ Our individual ability to collect data is limited
 - ▶ Better science
 - reproducibility
 - more vantage points
 - longitudinal view
- Data catalog doesn't instantly make for better science, but the process just might

Sharing Measurements (cont.)

- We share a little now
 - ▶ Mostly by people who are on a crusade, not random researchers
- Outlier: RouteViews
 - ▶ easy and useful
- Why don't we share more?
 - ▶ various reasons that we'll explore in an attempt to understand how we might change the situation
 - ▶ a cross-cutting key to keep in mind is scale

RouteViews

- Key aspects:
 - ▶ broad participation in data gathering
 - ▶ extremely useful to various sorts of people
 - ▶ longitudinal
- Drivers:
 - ▶ easy for operators to participate in
 - ▶ useful for operators and researchers
- Caveats:
 - ▶ only capturing one kind of data is easier than cataloging generic measurements

Practical Problems in Sharing

- Passive measurements (in particular) have many issues:
 - ▶ privacy/policy/legal hassles
 - ▶ competitive issues
 - ▶ lots of reasons to say "no", very little reason to say "yes"
- We should just accept that some data will never be released
- Active measurements, on the other hand ...

Laziness

- *Computer scientists are among the laziest people in the world*
 - ▶ it's part of our charm!
- It's generally just a time consuming hassle to cleanly package data to be released to others
 - ▶ especially for big datasets
 - ▶ e.g., active NIMI measurements used to validate a loss estimation scheme [AEO03]
 - ▶ we're not alone

Data Isn't Useful To Others

- It's not easy to package datasets in a way that is useful to others
 - ▶ because our measurements are stored as:
 - `run6/set12-32/netperf-100-6-06012004-foo.icir.org-...`
 - (in probes-nimi-7.tar.gz, of course!)
- Even if we could package this up and provide a README so that others could untangle our mess ...
 - ▶ we probably didn't collect the right meta-data to make the measurements generally useful
 - ▶ we didn't collect the context our measurements were conducted in (e.g., DNS to IP address mappings)
- I.e., we take measurements for our own purposes only

No Credit

- Researchers get no "credit" for releasing data
 - ▶ maybe an ACK in a paper
- Carefully gathering a dataset and keeping track of all the details is time consuming and worthwhile work
 - ▶ impact can be dramatic
 - ▶ effort is at least on par with writing a good piece of software
 - ▶ effort is at least on par with writing a good paper
 - ▶ *effort is much more involved than writing most of the papers / referee!*
- Not much funding for making datasets available
- Tenure boards don't care
- Management chains don't care

Cultural Shift

- **So, we need a cultural shift**
- We need researchers ...
 - ▶ to not be lazy
 - ▶ to collect meta data that serves no purpose for them, but very well could for others
 - ▶ to hold solid, public datasets in high-esteem
- A tall order ...

What Will It Take?

- Lots of *mundane work*
- As a community we need to commit to keeping a repository operational
 - ▶ not a small point
- A measurement repository must be easy and useful to researchers
 - ▶ e.g., RouteViews
- Measurement tools should help collect meta-data
 - ▶ e.g., `ipsu.mdump`
 - ▶ e.g., wrapper scripts

What Will It Take? (cont.)

- We need tools that help researchers integrate measurements into the catalog
 - ▶ i.e., if researchers have to fire up emacs and write a big block of XML for every measurement they take then it won't fly
- We need anonymization techniques that work
 - ▶ leave enough meat in the dataset
 - ▶ especially tricky for security measurements
 - e.g., ground truth datasets for stepping stone detection

A Plan

- Bump papers whose authors won't release the data
 - ▶ ok, maybe too drastic ...
 - ▶ but, changing the culture to one where it is expected that data is released
 - a tall order, but once the DC is in place we can start
- Concentrate on the easy stuff first: active measurements
- Find some pioneers to seed the system
 - ▶ you?
 - ▶ me?

A Plan (cont.)

- Make a "requirement" of mining the data from the system be to prominently ACK the system in papers
- Make data contribution a condition of funding (ala software in some cases)

A Plan (cont.)

- Or, some of your ideas