Trace anonymization misses the point

Jeffrey Mogul
Compaq? Western Research Lab
JeffMogul@acm.org

April 22, 2002

WWW 2002 panel on Web Measurements

What's wrong with anonymization?

Replication, generality $\Rightarrow$ need to use multiple traces

- Including corporations, ISPs, universities

Anonymization isn't good enough:

- Leaves too much stuff out:
  - Many experiments cannot use anonymized traces
- Leaves too much stuff in:
  - Many companies won't release anonymized traces because too much information remains
  - Anonymizer needs to be verified (I've seen errors!)

Solution: Move the code to the data

Return results, not anonymized inputs

Make code-shipping a well-understood practice:

- Set examples so that corporations, ISPs will accept it
  - Place code & results on Internet Traffic Archive
- Shipped code must be checkable $\Rightarrow$ use source code
  - Encourages checking & replication of experiments
- Would benefit from non-crappy standard log format
  - Don't want to reimplement code for each trace site

Use CLF, go to jail

Problems with Common Log Format and its relatives:

- Missing information (timestamps, durations, etc.)
- Lack of first-class support for proxies, caches
- Hard to parse
  - Crappy quoting mechanisms, timestamp formats

Needed: well-defined, highly-detailed Web log format

- Focus on external behavior, not server internals
- Should be extensible, self-describing

I didn’t say it would be easy!

Debugging shipped code is harder

- Usually requires a helper behind the curtain

Some ISPs, corporations will still resist

- Researchers must explain the benefits
- Will need some serious proofs of safety

Deployment of new log format will take time

- Incentives for commonality here are unclear
- Large sites more likely to be using their own code

Some stuff that I want in Web traces

- Timestamps
  - Fine-grained and synchronized
  - For all protocol events (e.g. conn. established, req. received, resp. starts, resp. ends, conn. closed)
- Cache-related headers:
  - Cache-control, Expires, Last-Modifed, Etag
  - If-Modified-Since, If-None-Match
- Connection IDs, encoding-related headers
- Accurate errors & sizes (including failed xfers)