On the Crucial Need for DITL Meta-Data

Mark Allman
International Computer Science Institute

DNS-OARC 35
May 2021

"I've seen a lot of sights,
And traveled many miles,
Shook a thousand hands
And seen my share of smiles…"

Thanks!

- Thanks to all the DITL contributors for making DNS data available!

- This talk is *not a complaint*, but a request for a little more help

Recording Reality

- Packet traces are an *approximation* of reality
  - we hope & strive for a *highly accurate approximation*
  - but stuff happens and deviations from reality are recorded …
    - … despite our wishes
    - … because of our wishes

Meta-Data Fills The Gaps

- Collecting a good set of meta-data can be crucially inform the analysis of data

Meta-Data Scope

- Lots of meta-data topics
  - will only focus on a few that impact DITL analysis to a large degree

Identify Missing Data

DITL Trace

Reality

Identify Missing Data

- We could have abridged data for a variety of reasons
  - e.g., tracing machine couldn't keep up at peak load
  - e.g., monitoring only incoming or outgoing link, but not both
  - e.g., measuring some, but not all replicas
  - e.g., monitor used packet sampling to keep up
  - e.g., the trace spans a maintenance window

Identify Transformed Data

DITL Trace

Reality

Identify Transformed Data

- When something in the trace is willfully changed, tell us!
  - e.g., IP addresses are obscured
  - e.g., some sensitive hostnames were obscured

Identify Transformed Data

- And, tell us *how* it was transformed
  - e.g., you used prefix-preserving IP obfuscation
  - e.g., you used random IP obfuscation
  - e.g., you applied the same (or different) transformation across all traces

Add Context

- Where data was captured
- Anything you can add about obvious puzzles would be great
  - e.g., drop out in packets for a while was real and caused by a routing issue
- Anything you can add that can't be readily distilled from the data is helpful
  - e.g., the one server IP address is really 10 servers behind a load balancer

Add Contact Information

- Researchers will find puzzles in the data
- If possible, providing an easy contact who knows about the data collection would be great

Providing Meta-Data

- Meta-data doesn't have to be fancy!
- First order issue is getting the meta-data to researchers
  - we'll cope with whatever you give us
  - E.g., a text file is fine!
  - Meta-data about existing datasets would be great, too!

What Can OARC Do?

- Accept non-pcap adjuncts to the packet traces
- Develop a template for data contributors to fill in asking for some common information

INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

Questions? Comments?

Mark Allman, *mallman@icir.org*
https://www.icir.org/mallman/
@mallman_icsi