Automatic Metadata Generation for Active Measurement

Joel Sommers
Colgate University

Ramakrishnan Durairajan
University of Oregon

Paul Barford
University of Wisconsin
comScore, Inc.

Active network measurement

- Long history of “probing the network” to measure phenomena of interest
  - IMC has seen its fair share of “clever” probing techniques!
- Lots of challenges to doing it *right*
  - Unexpected bias, noise, system effects
  - Do the measurements reflect the phenomenon under study? *How do we know?*

What does it mean to do active measurement well?

- **Sound** (Paxson, 2004)
  - Follow practices that lead to confidently saying that results are well-justified
  - Comprehensive metadata, measurement calibration, consider reproducibility, …
- **Hygienic** (Krishnamurthy, Willinger, Gill, and Arlitt, 2011)
  - Critical questions to ask in generating and consuming data during the research process
  - Producers: metadata to capture understanding about context of measurements; Consumers: use metadata to determine stretchability/appropriateness of data
- **Ethical** (Partridge and Allman, 2016)
  - Minimize risk of harm
  - Acceptable Use Policy should accompany release of data (Allman and Paxson, 2007)

Common element of advice: capture metadata

- Data can (and do!) live beyond a given study
- Metadata: some representation of the original context of experiments
  - Use to understand precision, data format, measurement process, options used with tools, software versions, etc.
  - Use to understand quality and scope of measurements (and results)
  - Use to understand how to replicate or reproduce results
  - Potential for evaluating and correcting for measurement bias

Goal of this work

- Create a tool for a*utomatic* capture of metadata for active measurement experiments
  - Make it easy and thus possibly routine
  - Make it extensible
    - Simple metadata analysis scripts
    - Cross-platform, lightweight
- We take a broad view of metadata
  - Descriptive information (similar to existing/prior data collection efforts)
  - Periodic system measures, e.g., CPU load, network load, memory activity
  - Periodic measurement of local network conditions, e.g., RTTs to first k hops

https://xkcd.com/917/

SoMeta: the xkcd

SoMeta: the tool

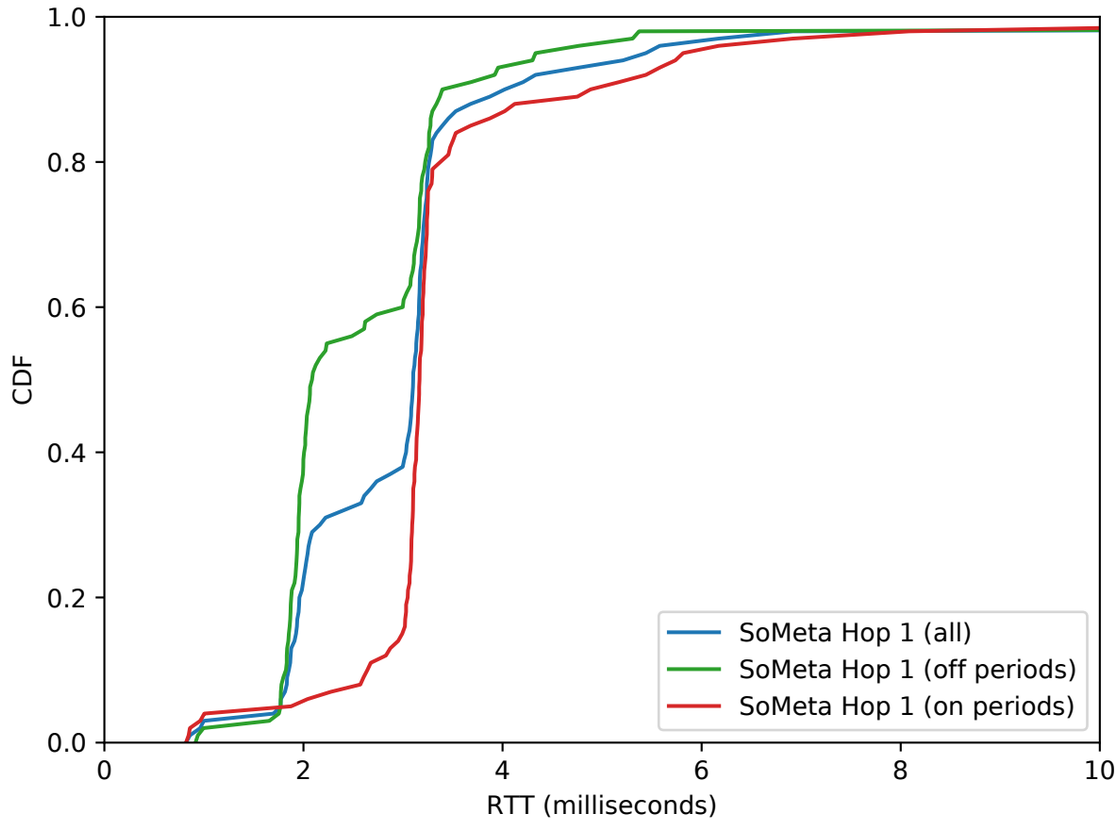
- Written in Python, the basis of another famous xkcd (https://xkcd.com/353/)
  - Scheduling engine based on asyncio and coroutines; debugging can be totally meta
  - The measurement tool is started as a subprocess of SoMeta
    - Provides seamless integration of existing measurement tools
- Includes a set of configurable & extensible monitors
  - CPU, IO, memory, netstat (based on psutil); not meta, but still nice
  - RTT monitor; can emit ICMP/UDP/TCP probes
    - Uses Switchyard [Sommers 2015] to access libpcap and parse/construct packets (which is quite meta, if I might say so)
- Collected metadata written to a file in a self-describing JSON format, thus meta
- Includes basic tools for meta-analysis, e.g., plotting and summary statistics, so meta^2

SoMeta Evaluation

- Experiments designed to
  - Evaluate overheads of SoMeta on different systems, e.g., Raspberry Pi (v1 and v3), mid-range server
    - Measure CPU, memory, I/O, network, RTT every 1s or every 5s
      - Also used monitors individually at different measurement intervals
  - Illustrate how metadata might be used to identify poor/biased measurements due to system and local-network effects
    - Artificial CPU, memory, I/O, and network loads in a lab environment and in a home broadband environment
    - Used Scamper as the active measurement tool (ping mode)

Results summary

- Overhead experiments:
  - On lowest-end Pi 1 model B of 12% avg CPU
  - 1-3% on a Pi 3 model B, <1% on server
- Artificial load experiments:
  - SoMeta system measurements reveal high load
  - SoMeta RTT measurements to local targets reveal clear shift in RTTs, also present in the Scamper measurements



SoMeta RTTs gathered from artificial CPU load experiment (using a Pi 3)

Summary and ongoing work

- SoMeta is a tool to help automate collection of metadata from active measurement experiments
    - Goal: address the long-standing need to capture more comprehensive contextual information about active measurement experiments
- Future/ongoing work
    - Development of standard configurations and best practices for metadata collection and publication
    - More comprehensive post-processing and analysis of metadata
    - Modifications for lightweight metadata capture in multi-user/multi-measurement environments
- Code: https://github.com/jsommers/metameasurement
    - Figures in paper are clickable and lead to results generation scripts on github

Thanks!

https://github.com/jsommers/metameasurement
jsommers@colgate.edu