Computation and Aggregation of Quantiles from Data Streams

John Chambers, David James,
Diane Lambert, Scott Vander Wiel

Vienna, June 17, 2006

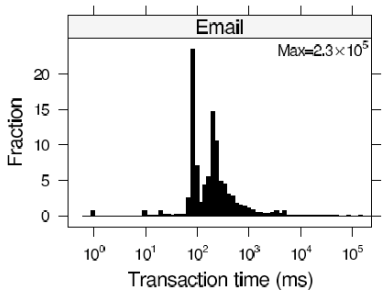
(related article to appear with discussion in "Statistical Science")

Motivation

- Application at Lucent Technologies: software to monitor distributed IP-based services.
- Goal: characterize various metrics (e.g. e-mail transaction times), locally and aggregated, updated over time.
- Constraint: computing at the node, amount of data transmitted to server.

Quantile Estimation

Metrics are often unusually distributed (long tails, bimodal, ...)



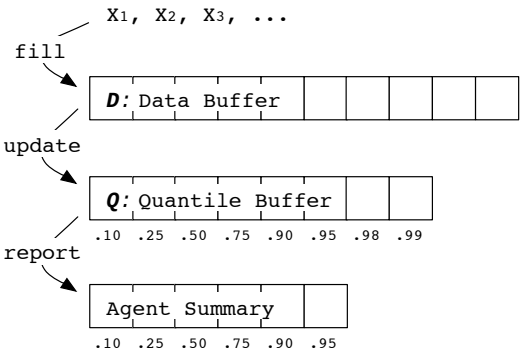
Need to estimate quantiles (often in tail).

The Idea

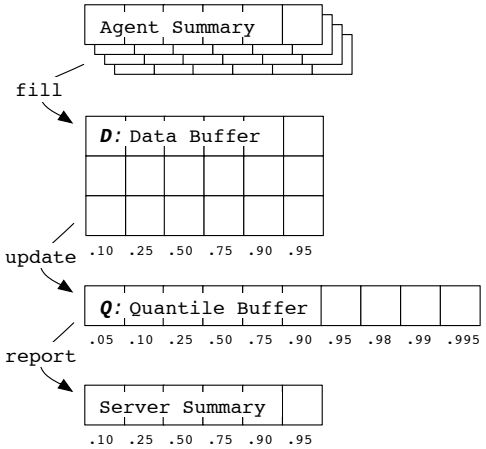
(Approximate, Update, Aggregate)

- Approximate the empirical distribution for each metric & node (agent)
- Update each approximation periodically for new data at the node.
- Aggregate the ecdfs for relevant groupings of nodes (e.g., regions)

Update for each agent



Aggregate agent records



Software

- Objects represent each evolving quantile estimate: `a <- seqQuants(....)`
- OOP-style functions to simulate updating, aggregating: `a$merge(data)` (modifies `a`)
- Using R closures (object contains functions with a shared environment for updates).

Software

- R simplifies large-scale simulation studies, with varying statistical assumptions.
- R also helps in the algorithm development in C, by calling an R tracer from C.

Summary

- An example of the productive interaction between applications and research, typical of Bell Labs research (in the old days).
- An interesting algorithmic study to estimate distributions with distributed, ongoing data.
- The productive computing environment centered on R essential for productivity.