

Do Web Measurements Measure Up?



Mark Crovella

Boston University Computer Science

Some Observations

- Web measurement efforts have struggled with
 - Size: there is too much to measure
 - Complexity: Protocols and Systems
 - HTTP/1.0 vs. HTTP/1.1
 - Proxies, Server Clusters, CDNs
 - Change: workloads have changed faster than we can keep up
 - “Web” is an umbrella for an evolving information system
- It's time to take stock
 - We can't measure everything we'd like
 - What should our strategy be?

All Web Measurements are Incomplete

... but some are useful

What should our measurement strategy be?

What distinguishes useful from useless measurements?

Why are we Measuring?

- We want to design new systems and improve existing systems
 - Need to focus on measurements that can guide designs
 - Timescale of designs is much longer than timescale of change
- How can we do this?
 - Drive design: focus on metrics that matter
 - Fight change: focus on invariants

Where is the insight?

Example: Cache replacement policies

A vast number of studies have looked at cache replacement

From a design perspective, little has been learned beyond:

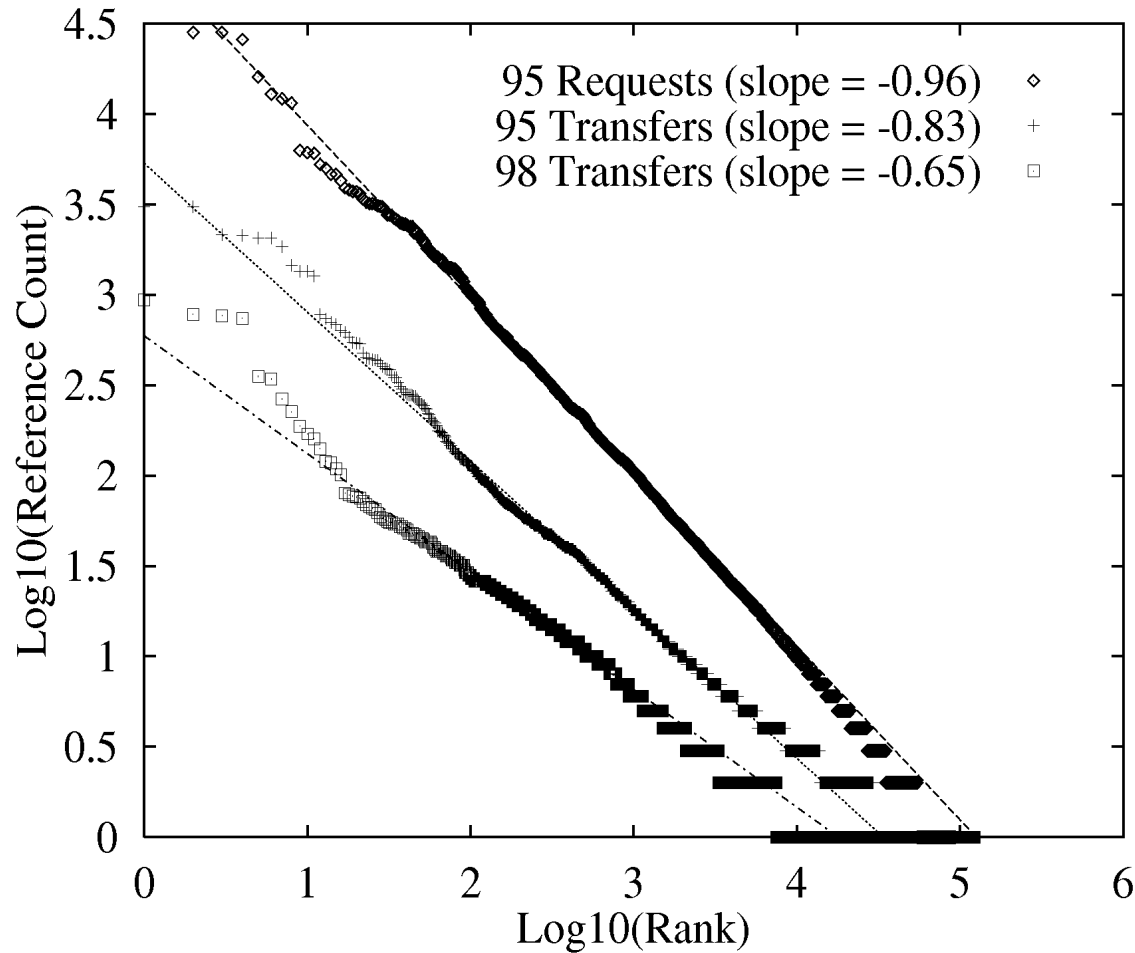
- Caches will never solve problems of latency and bandwidth
- Cache hit rates will never approach 100%

The important question is: *Why?*

Where are the theories?

- In processor memory caches, hit rates regularly achieve $> 99\%$
- In the Web, poor cache effectiveness seems to be due to
 - Zipf's Law
 - Heavy tails

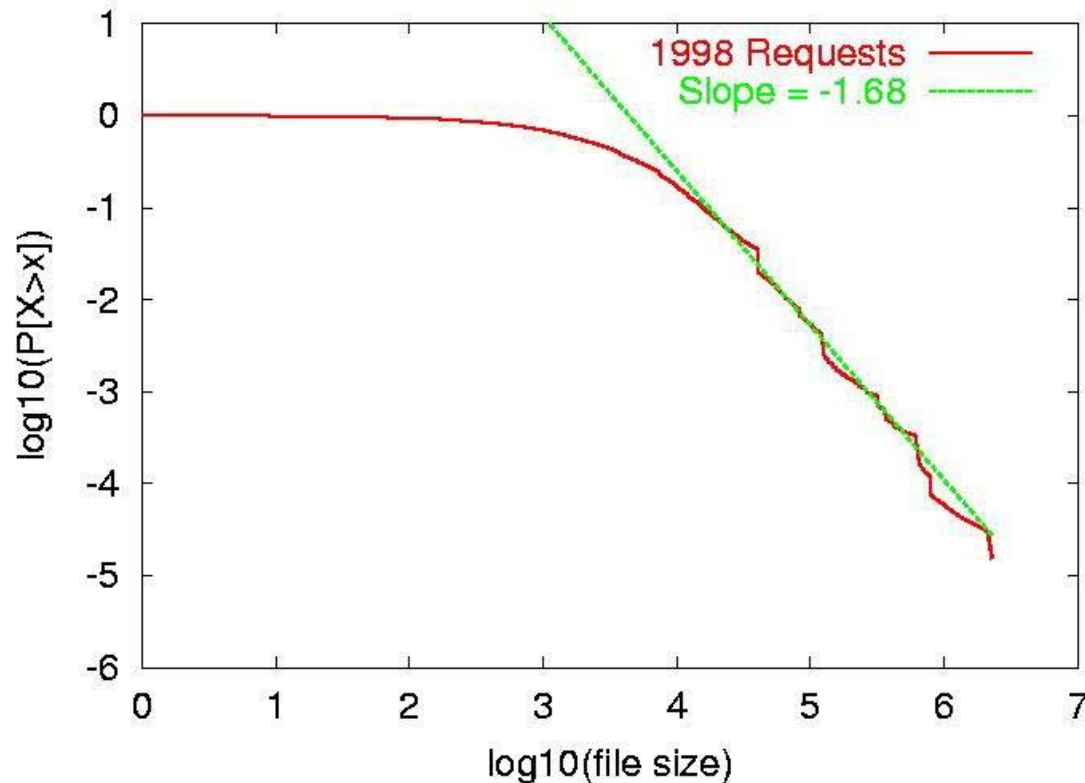
An invariant: Zipf's Law



Why do
object requests
follow
Zipf's Law?

We don't know.

Another invariant: heavy tails



Why are
object sizes
heavy-tailed?

We don't know.

$$P[X > x] \sim x^{-\alpha} \quad 0 < \alpha \leq 2$$

Science-Driven Measurement

- Thesis: instead of studying cache hit rates, we should be studying Zipf laws and heavy tails.
 - Why they arise
 - What influences their properties
- Theories last past the timescale over which measurements change
- We need a focus on "basic" research, instead of
 - Churning out numbers
 - Obsessing over short timescales
 - Inventing metrics