Rate of Change in the Web

Fred Douglis

AT&T Labs - Research

Joint work with Anja Feldmann, Balachander Krishnamurthy, and Jeffrey Mogul

Based on presentation at USITS'97

Overview

- Who cares about changing data?
  - Caches
    - Update cached copies upon change
    - Detect pages that should never be cached
    - Predict when to revalidate cached page
  - Users
    - Find out when to look at a changed page of importance
- Empirical study of the rate of change of Web data
  - Rate of reaccess to unchanged and changed resources
    - Able to detect certain regularities
  - Unfortunately, the dataset is dated by now

Metrics

- Frequency of reaccess
  - Directly affects cacheability
- Rate of change (fraction of changed accesses)
  - Affects both cacheability and delta-encoding
- Age of resources and modification intervals
  - Detect patterns
  - When to expire cached copies?
- Duplication of content
  - Mirrors, aliases
- Semantic differences
  - Content-specific changes

ROC: Contributing Factors

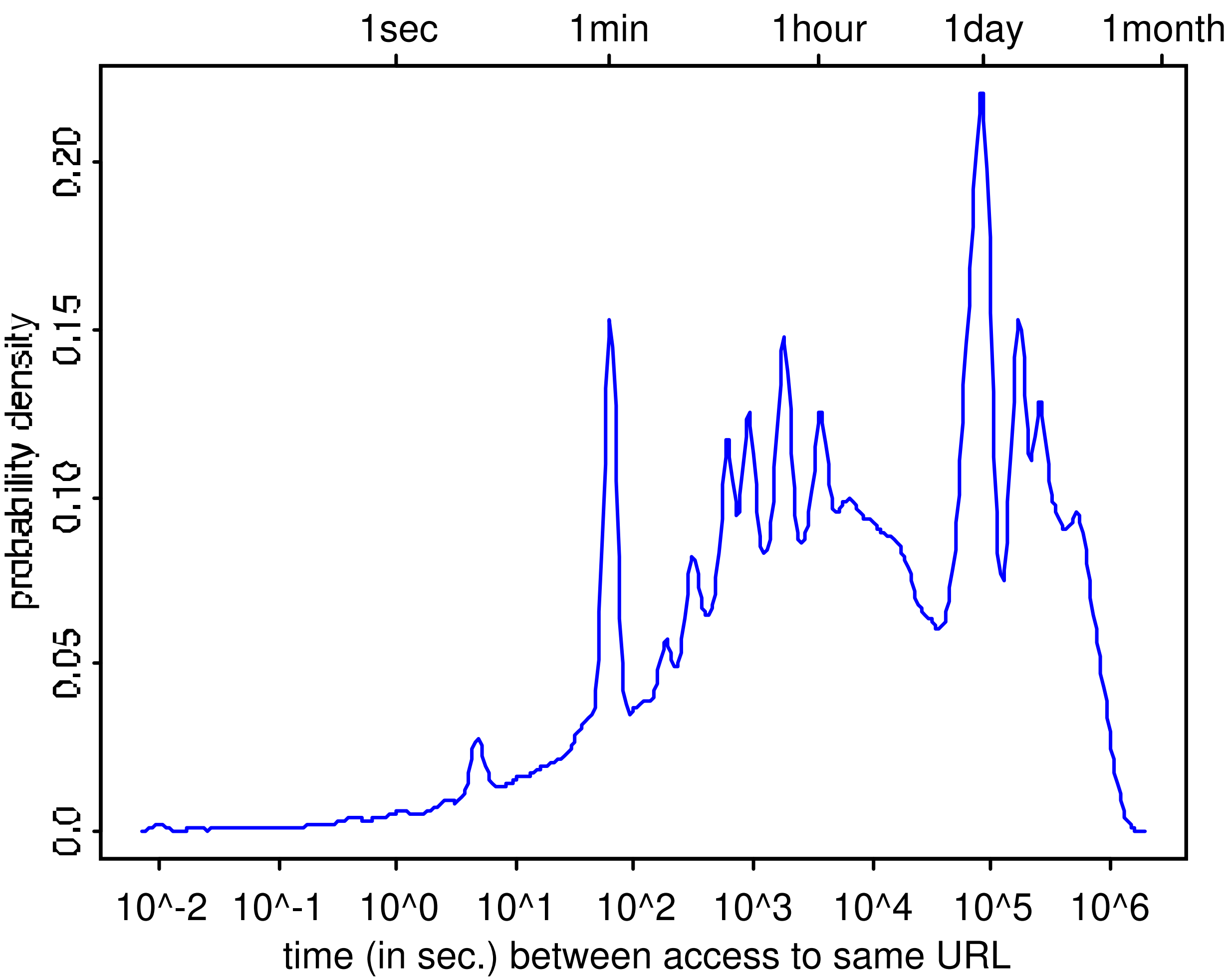
- Frequency of access
  - More frequently accessed resources change more often
- Content type
  - Images more static
- Size
  - Relatively unimportant
- Content-provider
  - Commercial sites have most dynamic content

Methodology

- Collected trace of outgoing HTTP traffic
  - All packets sent to port 80 outside AT&T Labs -- Research
    - November 1996 dataset
  - Reconstructed full content of all requests and responses
  - 17 days' activity, 465 clients, 21K servers
  - See Anja's talk
- Second trace, from DECWRL, with similar results
  - But just 2 days, and omitted nontextual resources
  - Not discussed here

Per-URL Access Rate

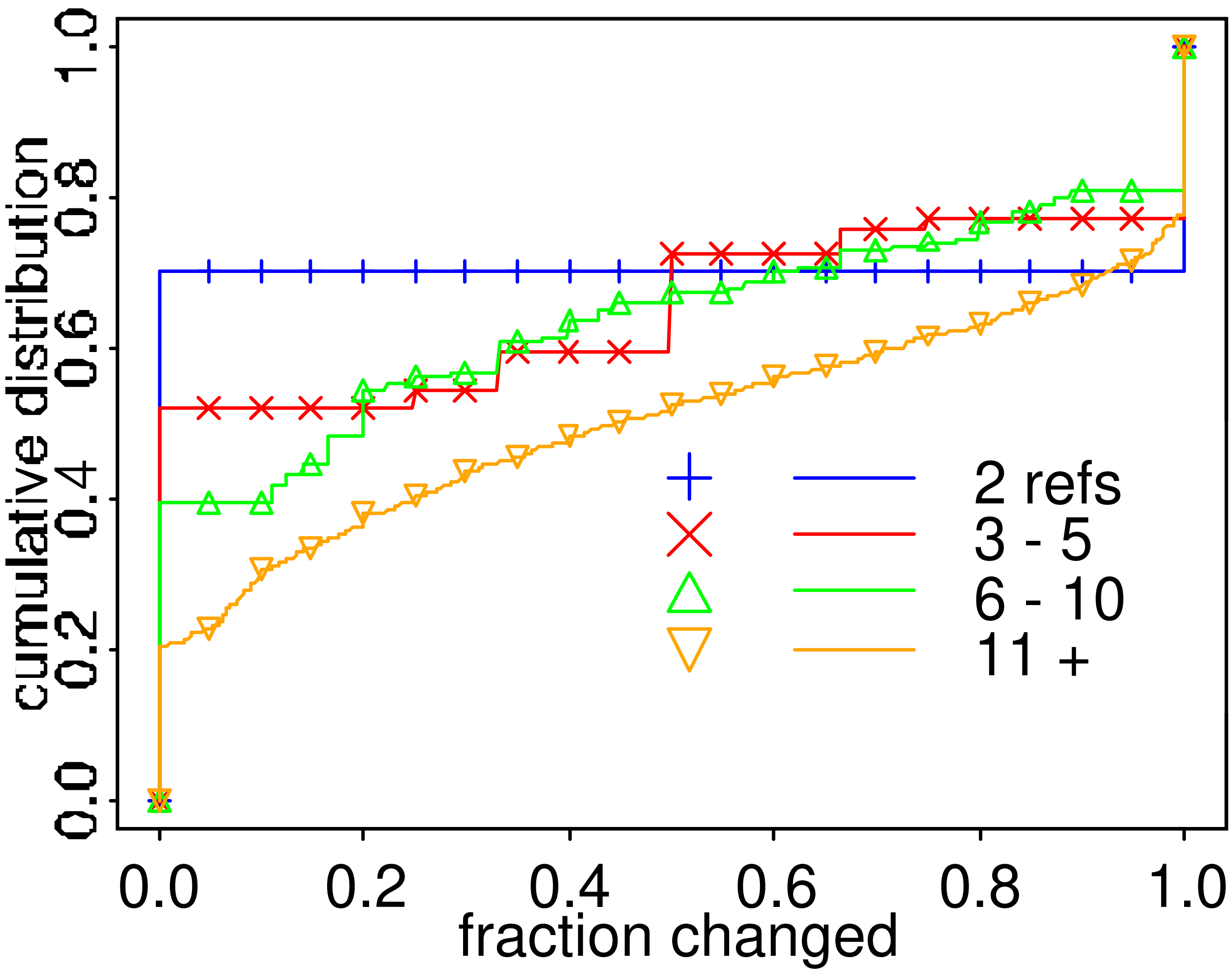
- Pronounced intervals of access
  - Mean 25.4 hrs
  - Median 1.9 hrs
- Reaccess rate:
  - resources: 22%
  - accesses: 49%



Change Ratio

- What fraction of accesses are to changed data?
  - Images rarely change
  - Octet-stream nearly always changes
  - HTML bimodal
    - About half never change
    - Most of the rest change on each access
- Resources accessed just 2-3 times can <u>skew</u> change ratio
  - 100% change ratio if seen twice and second access changed
- Consider HTML resources by frequency of access

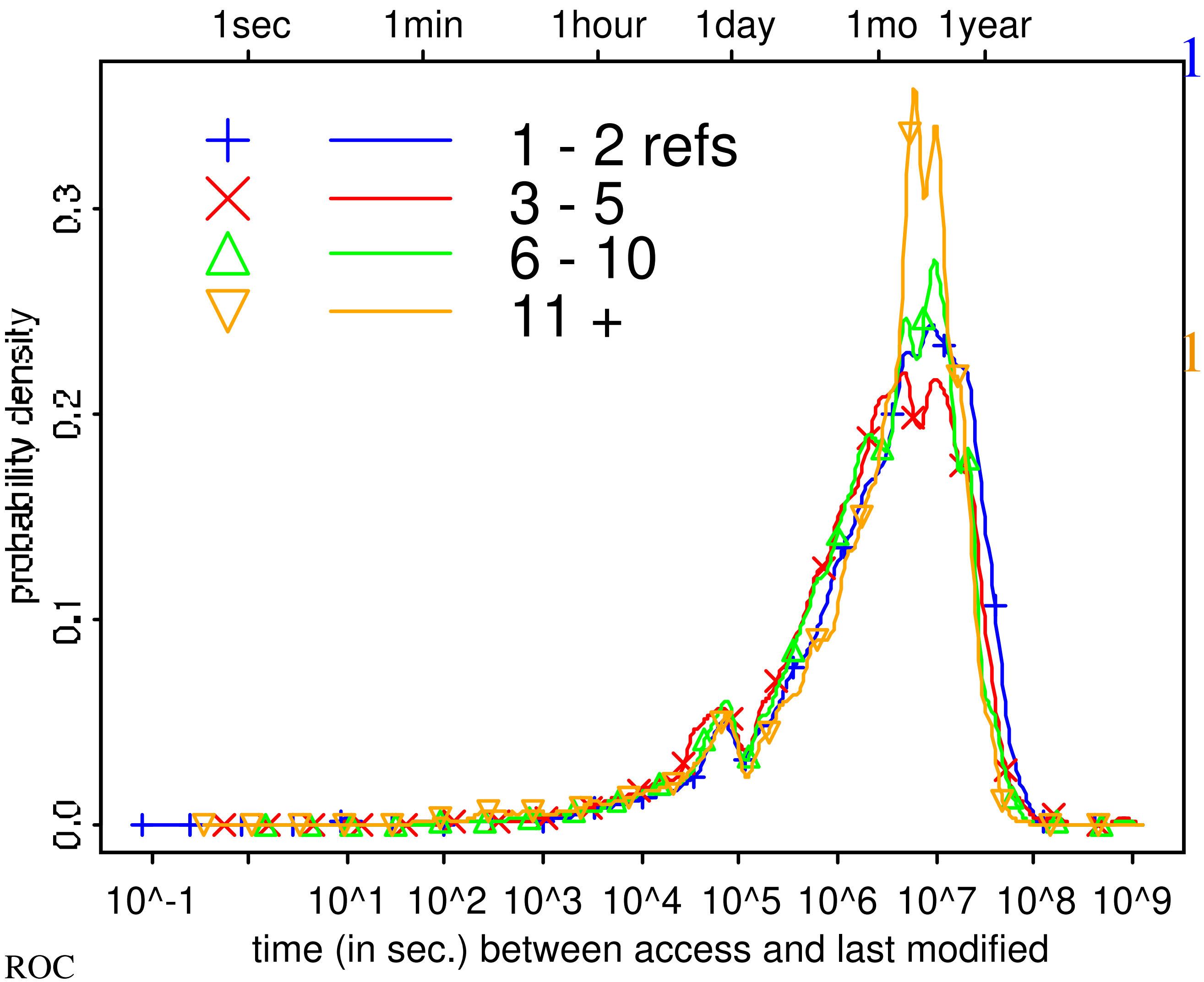
Change Ratio (HTML)



Age of Resources

- Excluding dynamic data, how old are resources?
  - Considered all responses with Last-Modified timestamps
- Number of references:
  - Most frequently accessed resources are younger
- Content type:
  - HTML younger than GIF
- Size:
  - Smaller (images) tend to be older
- Top-level domain (e.g. edu, com):
  - Educational sites are older

Age by Access Rate



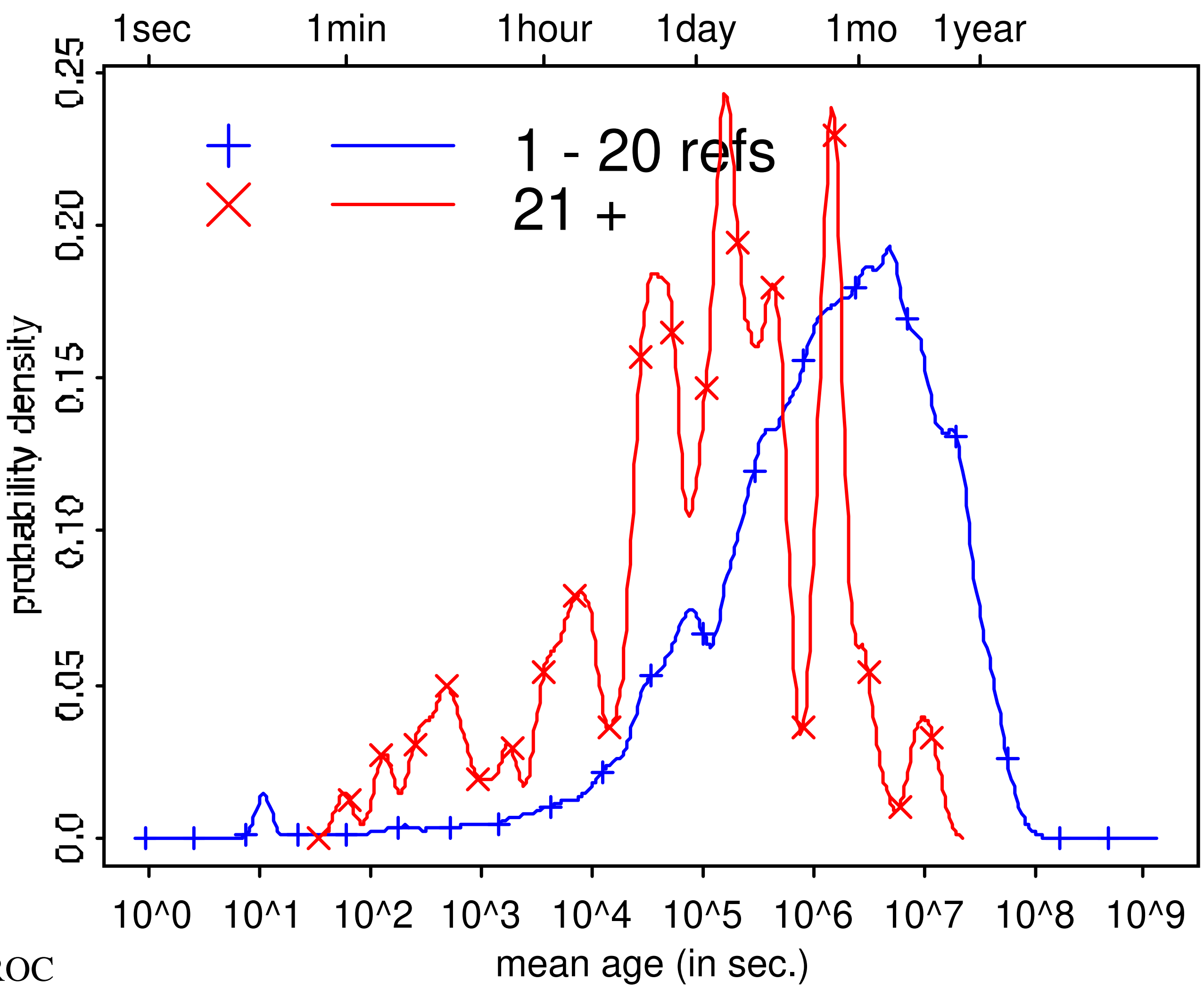
1-2 refs

- Mean: 120 days
- Median: 53
- Std Dev: 331

11+ refs

- Mean: 81 days
- Median: 54
- Std Dev: 157

Age by Access Rate (HTML)



1-20 refs

- Mean: 68 days
- Median: 20
- Std Dev: 140

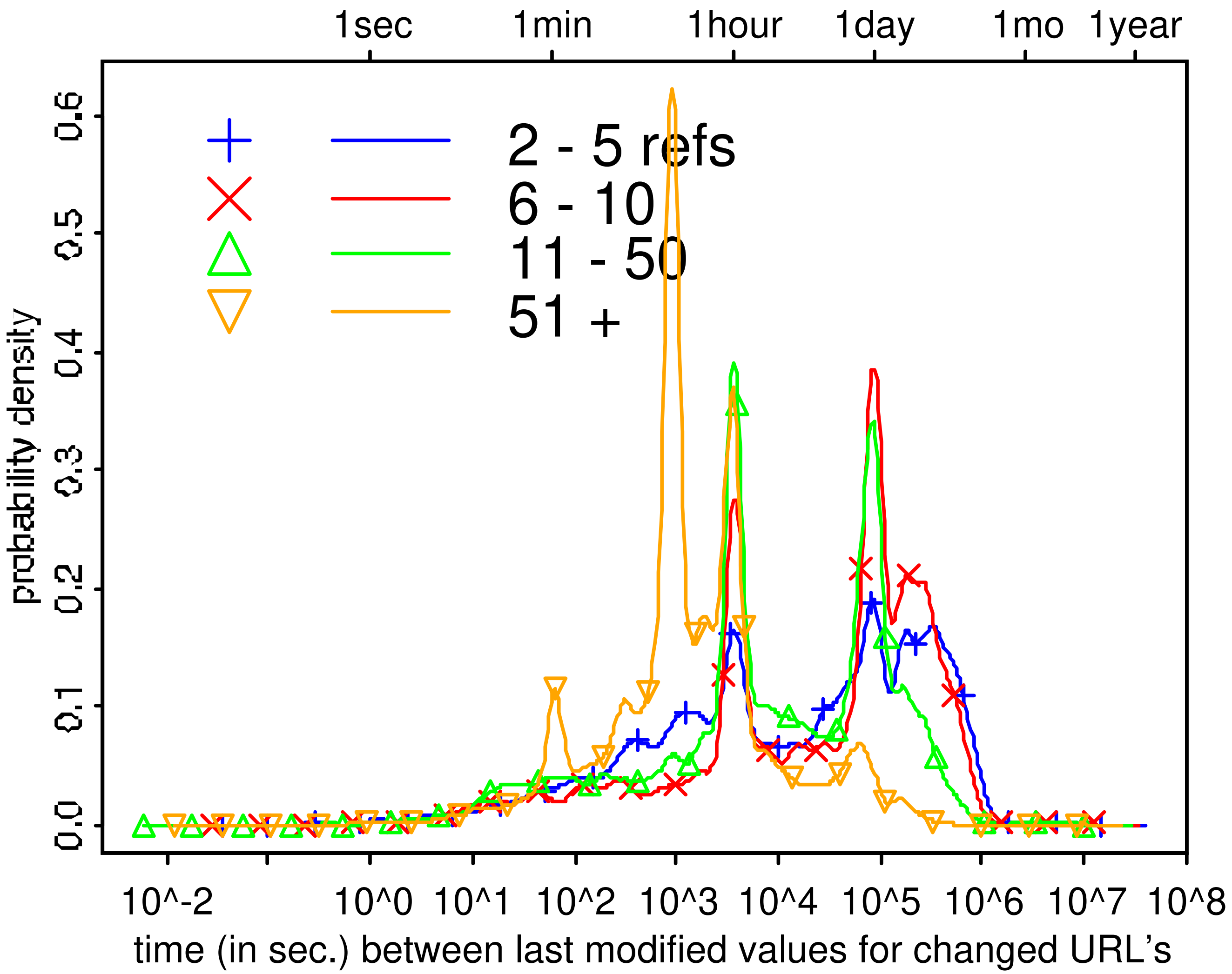
21+ refs

- Mean: 7.5 days
- Median: 1.5
- Std Dev: 20

Modification Intervals

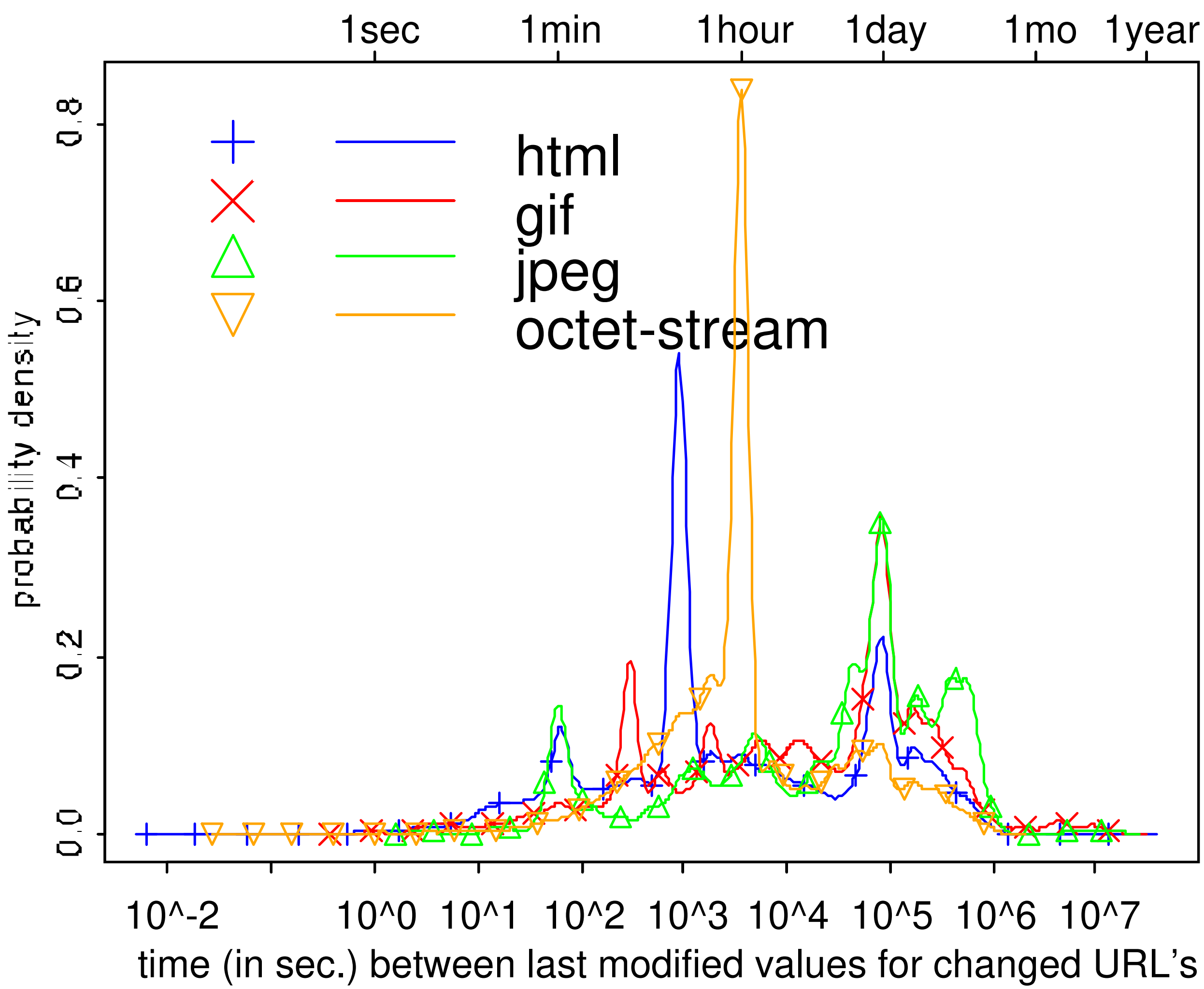
- Consider the difference between last-modified timestamps on successive accesses to changed resources
  - Use the time of the request if none available
- Peaks at certain intervals representing periodic updates
  - E.g., tickers, daily news

Modification Intervals (Access)



Modification Intervals (Content)

- Octet-stream accesses 1hr apart
- HTML tickers with 15m update intervals
- Daily updates to images



How to Use this Data?

- Can we tune heuristics?
  - When to expire cache?
  - When to poll a resource for URL-minder/AIDE?
- Will multicasting updates work?
- Should delta-encoding become a standard?

Conclusions

- Many resources change frequently, and these account for most accesses
  - Delta-encoding may prove more useful than simple caching
  - Many factors are involved
- Repeating these studies on current datasets might be interesting
- An infrastructure for propagating updates should prove useful
  - Eaglewatch is an example of this, but relies on cooperation from the content provider