Deterministic Algorithms for Biased Quantiles

Graham Cormode
cormode@bell-labs.com

Flip Korn
flip@research.att.com

S. Muthukrishnan
muthu@cs.rutgers.edu

Divesh Srivastava
divesh@research.att.com

Quantiles

Quantiles summarize data distribution concisely.

Given N items, the ϕ–quantile is the item with *rank* ϕN in the sorted order.

Eg. The median is the 0.5-quantile, the minimum is the 0-quantile.

Equidepth histograms put bucket boundaries on regular quantile values, eg 0.1, 0.2…0.9

Quantiles are a robust and rich summary: median is less affected by outliers than mean

Quantiles over Data Streams

Data stream consists items arriving in arbitrary order, number (so far) is N.

Models many data sources eg network traffic, each packet is one item.

Requires linear space to compute quantiles exactly in one pass, $\Omega(N^{1/p})$ in p passes [MP80].

ε-approximate computation in sub-linear space

- Φ-quantile: item with rank between $(\Phi-\varepsilon)N$ and $(\Phi+\varepsilon)N$
- [GK01]: insertions only, space $O(1/\varepsilon \log(\varepsilon N))$
- [CM04]: insertions & deletions, space $O(1/\varepsilon \log^2 U \log 1/\delta)$

Why Biased Quantiles?

IP network traffic is very skewed

- Long tails of great interest
- Eg: 0.9, 0.95, 0.99-quantiles of TCP round trip times

Issue: uniform error guarantees

- ε = 0.05: okay for median, but not 0.99-quantile
- ε = 0.001: okay for both, but needs too much space

Goal: support *relative* error guarantees in small space

- Low-biased quantiles: ϕ-quantiles in ranks $\phi(1\pm\varepsilon)N$
- High-biased quantiles: $(1-\phi)$-quantiles in ranks $(1-(1\pm\varepsilon)\phi)N$

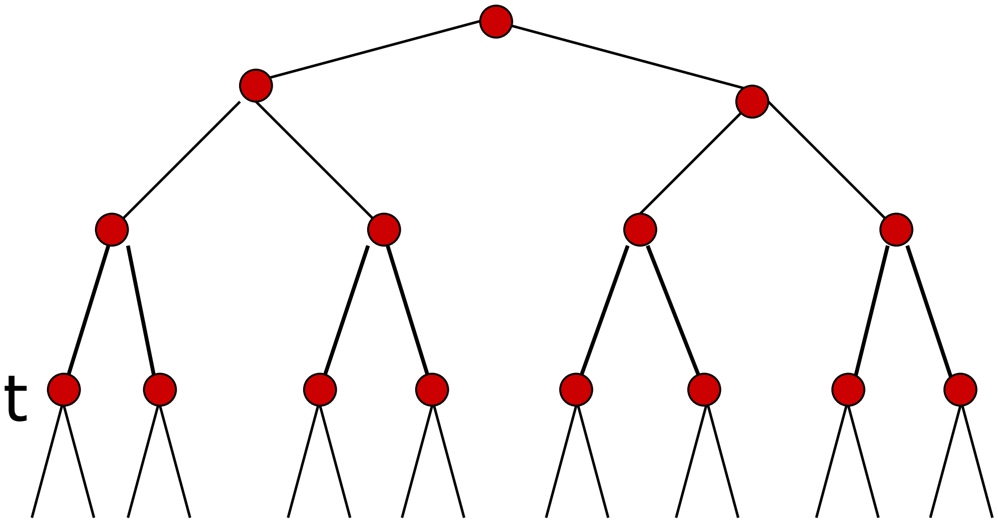
Prior Work

- Sampling approach due to Gupta & Zane [GZ03]
  - Keep $O(1/\varepsilon \log N)$ samplers at different sample rates, each keeping a sample of $O(1/\varepsilon^2)$ items
  - Total space: $O(1/\varepsilon^3)$, probabilistic algorithm
- Deterministic alg [CKMS05]
  - Worst case input causes linear space usage
  - Showed lower bound of $\Omega(1/\varepsilon \log \varepsilon N)$ for any alg
- Improved probabilistic alg of Zhang+ [ZLXKW06]
  - Needs $O(1/\varepsilon^2 \text{ polylog } N)$ space and time

Our Approach

Domain-oriented approach: items drawn from [1…U], want space to depend on O(log U)

- Impose binary tree structure over domain
- Maintain counts c_w on (subset of) nodes
- Count represents input items from that subtree



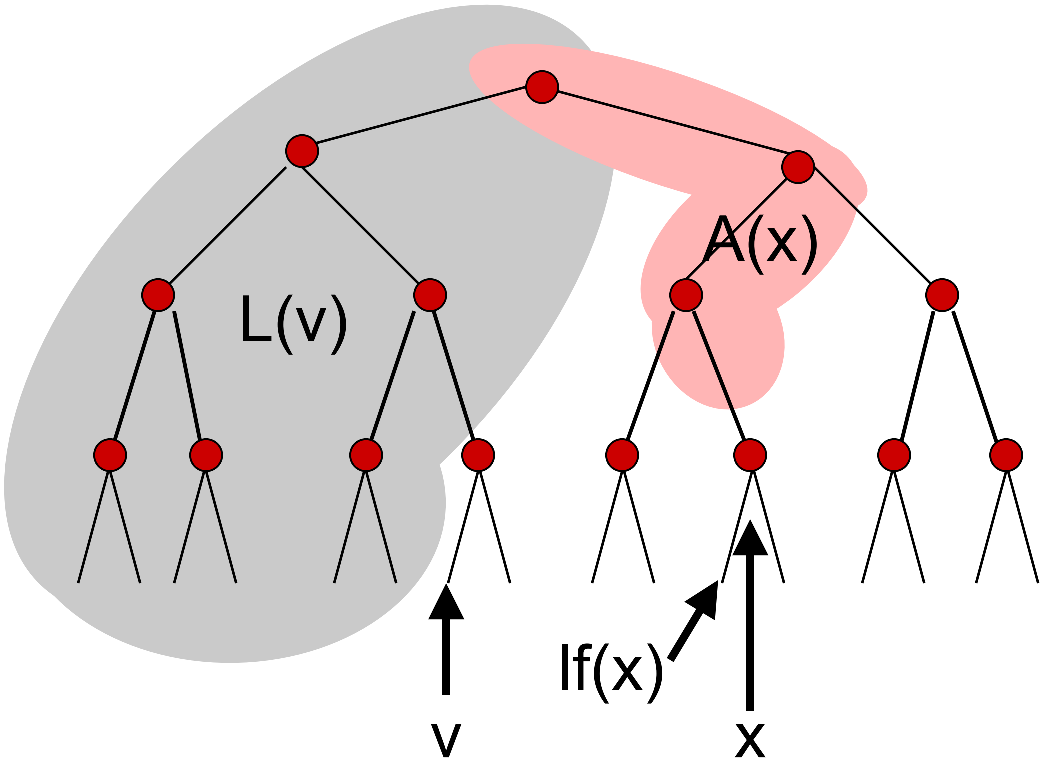
So counts to left of a leaf are from items strictly less; uncertainty in rank of item is from ancestors

Similar to [SBAS04] approach for uniform quantiles

Functions over the tree

We define some functions to measure counts over the tree.

- lf(x) = leftmost leaf in subtree x
- anc(x) = set of ancestors of node x
- $L(v) = \sum_{lf(w) < lf(v)} c_w$ (Left count)
- $A(x) = \sum_{w \in anc(x)} c_w$ (Ancestor count)



Accuracy Invariants

To ensure accurate answers, we maintain two invariants over the set of counts:

$$\forall x.\ L(x) - A(x) \leq rank(x) \leq L(x) \qquad ❶$$

ensures we can deterministically bound ranks

$$\forall v.\ v \neq lf(v) \Rightarrow (c_v \leq \alpha\, L(v)) \qquad ❷$$

ensures range of possible ranks is bounded

To guarantee ε-accurate ranks, will set $\alpha = \varepsilon/\log U$ (since we use ❷ summed over log U ancestors)

Claim: any summary satisfying ❶ and ❷ allows us to find $r'(x)$ so $|r'(x) - rank(x)| \leq \varepsilon\, rank(x)$

Data Structure

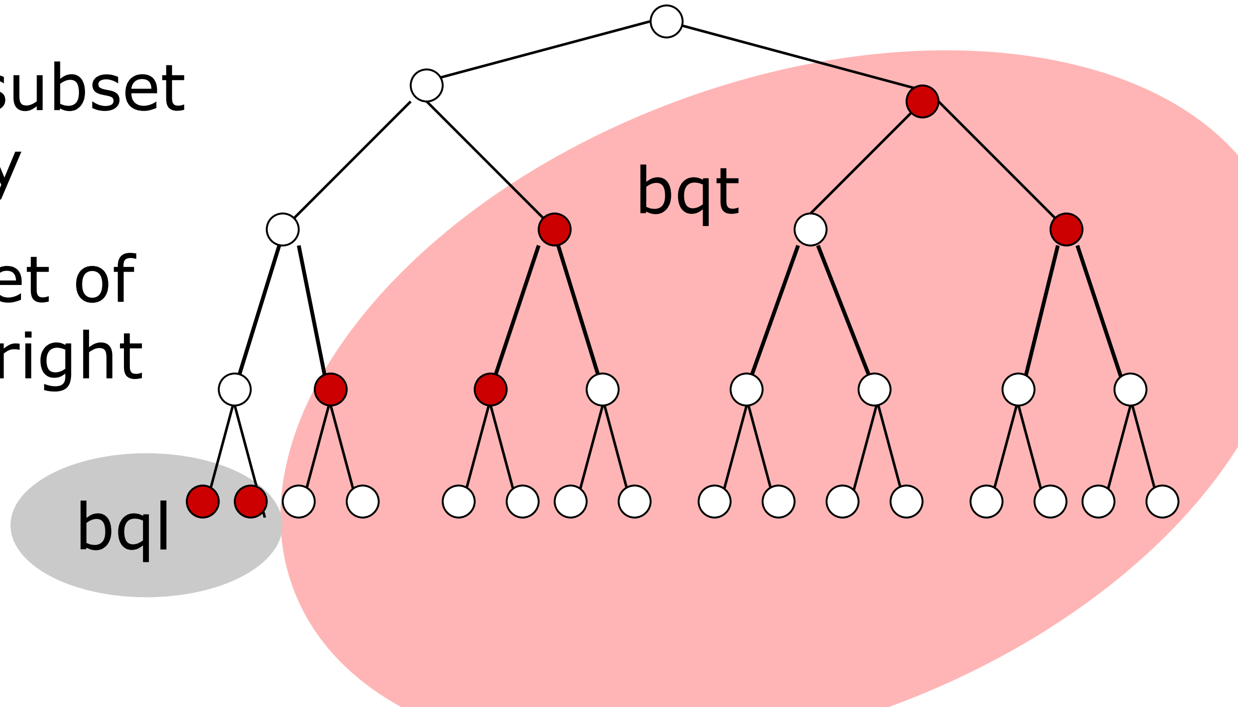
Store subset of nodes and counts as "bq-summary"

Nodes with count 0 do not need to be stored

Split bq into two: bq-leaves (bql) and bq-tree (bqt).

This division is needed to get tightest space bounds.

- bq-leaves is a subset of leaf nodes only
- bq-tree is subset of nodes strictly to right of bq-leaves



Equivalence Classes

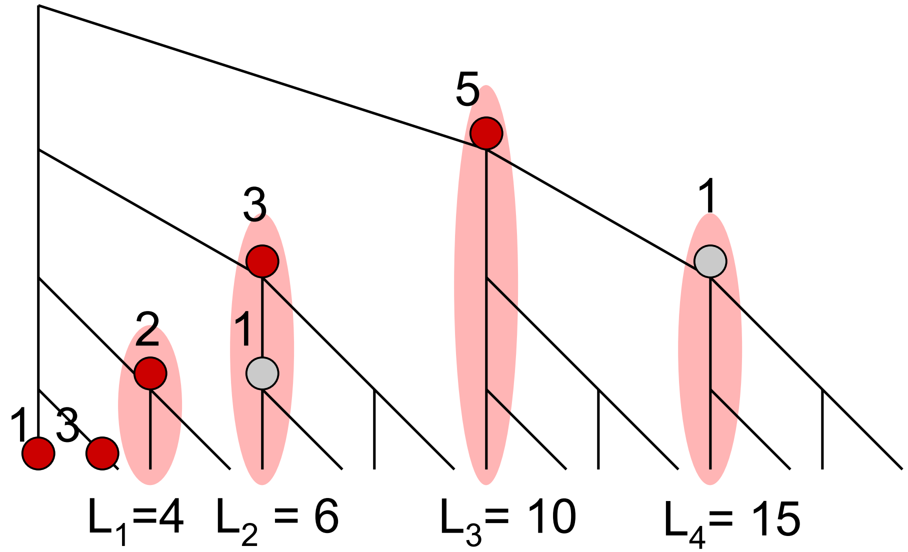
Main effort for the space bound is in proving that the size of bqt is bounded

We force all nodes V in bqt with at least one child present) to be "full": for $v \in V$, $c_v = \alpha L(v)$

- Partition V into equivalence classes based on $L(v)$: classes form paths

- E_i is set of nodes in i'th equivalence class, with L value = L_i

- L_1 is sum of bq-leaves: $L_1 = \sum_{v \in bql} c_v$



Example with $\alpha = ½$

Space Bound

- Ensure the number of leaves $|bql| = L_1 \geq 1/\alpha$
- The L_i's increase exponentially, can show $L_{i+1} \geq L_1 \Pi_{j=1}^{i}(1+\alpha|E_j|)$
  - Consider item $U+1$, so $rank(U+1)=N$.
  - Also $N = L(U+1) \geq 1/\alpha \, \Pi_{j=1}^{q} (1 + \alpha|E_j|)$
- Using these expressions, we bound size of $|bqt|$
- Total space $= |bql| + |bqt|$

$$\boxed{= O(1/\varepsilon \log (\varepsilon N) \log U)}$$

Maintenance

Need to show how to maintain the accuracy invariants, while guaranteeing space is bounded and updates are fast.

- Will `Insert` each update x. `Insert` will be defined to maintain accuracy, but space may grow
- Periodically will run a linear scan of data structure to `Compress` it.
- Will argue that these two together maintain space and time bounds.

Insert Procedure

Given update item x:

- Compare to $z = \max_{u \in bql} u$
- If $x \leq z$, place x in bql in time $O(1)$
- If $x > z$ place x in bqt in time $O(\log \log U)$:
  - Find closest materialized ancestor y of x in bqt
  - Add 1 to c_y unless this would make $c_y > \alpha L(y)$, if so then create child of y with count = 1

Easy to show this procedure maintains accuracy invariants. Space increases by ≤ 1 node.

Compress

- If we keep `Insert`ing, space can grow without limit, but in worst case, we add one new node per insert, so `Compress` when space doubles
- Need to periodically recompute L() values for nodes, and merge together nodes when possible
  - First, resize bq-leaves so $|bql| = \min(N, 1/\alpha)$
  - Recompute $z = \max_{v \in bql} v$ in time linear in $|bql|$, `Insert` leaves removed from $|bql|$ into bqt.
  - Tricky part is compressing bq-tree…

Compress Tree

- `CompressTree` operation takes a (sub)tree in bqt, ensures that each node is "full" (has $c_v = \alpha L(v)$) by "pulling up" weight from below
  - For node v compute $L(v)$ and $wt(v) = \sum_{v \in anc(w)} c_w$
  - Set c_v as big as possible by borrowing from $wt(v)$
  - Allows us to remove descendents with zero count
- With care, `CompressTree` takes time $O(|bqt|)$ and computes $L(v)$ incrementally as a side effect
- Can show `Compress` maintains conditions ❶, ❷, and the space bound follows.

Final Result

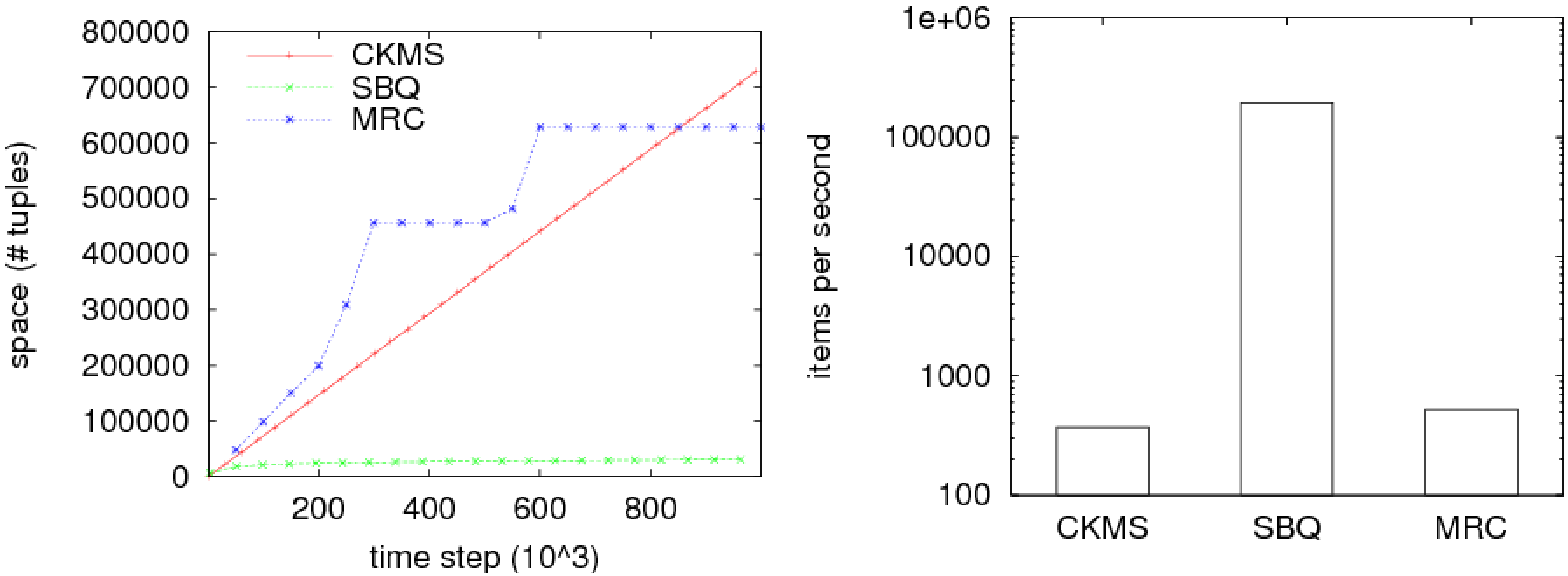
- Can answer rank queries with error ε rank(x), using space $O(1/\varepsilon \log \varepsilon N \log U)$, and amortized update time $O(\log \log U)$.
  - Lower bound on space = $O(1/\varepsilon \log (\varepsilon N))$
- To answer queries, need latest values of $L(v)$, so need time $O(1/\varepsilon \log \varepsilon N \log U)$ to preprocess
  - Can then answer queries in time $O(\log U)$ each
  - Alternatively, spend $O(\log U)$ time on updates and allow $L(v)$ values to be computed in time $O(\log U)$
  - Quantile queries can be answered by binary searching for item with desired rank

Extensions

- Partially biased algorithm
  - Sometimes only need accuracy down to some $\varepsilon' N$
  - Can reduce space slightly for this weaker guarantee
  - Space required is $O(1/\varepsilon \log (\varepsilon/\varepsilon') \log U)$
- Uniform algorithm
  - The `CompressTree` idea can be applied to εN error
  - bq-leaves not needed, space used is $O(1/\varepsilon \log U)$
  - Time is $O(\log \log U)$ amortized as before

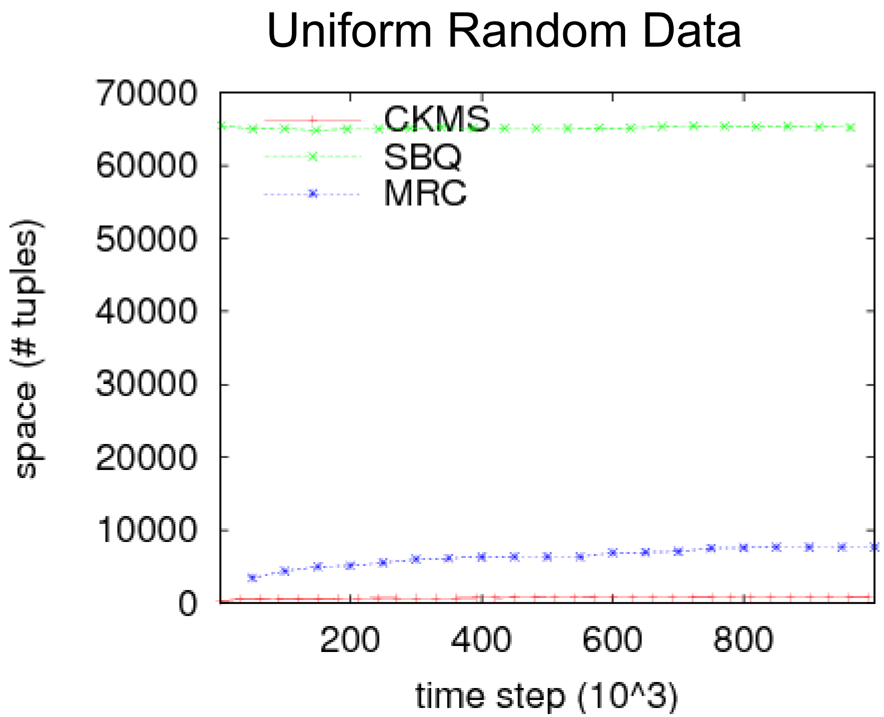
Experimental Results

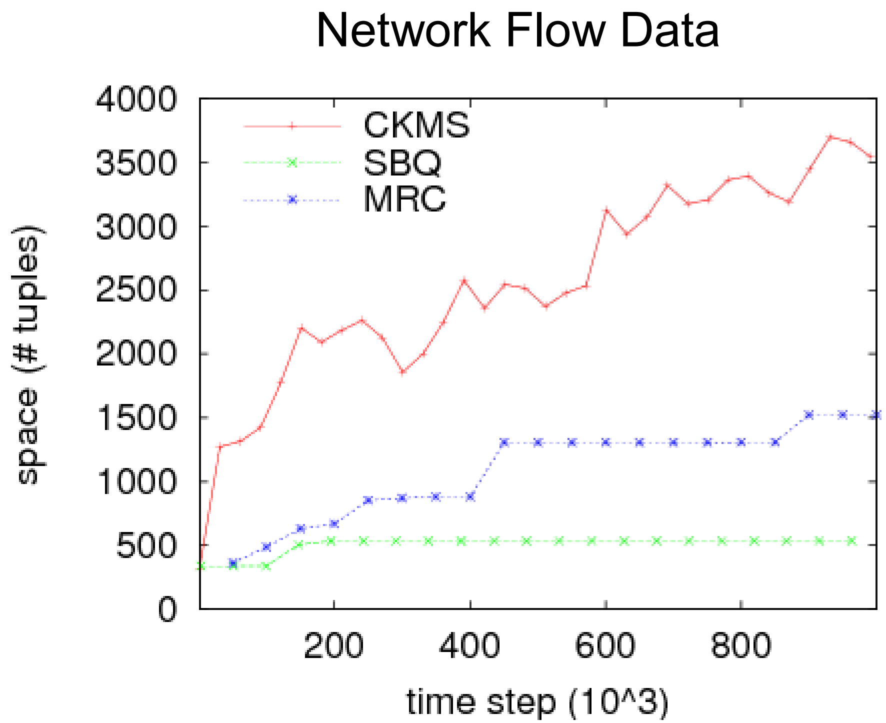
Nearly Sorted (worst case) data



- CKMS, MRC = prior work, SBQ = this work
- SBC has better space on some data sets
- SBC at least 25x faster than MRC on all data sets

Experimental results





- New alg can use more space than existing algs,
- Total space still small (in absolute terms)

Commentary

- Took effort to get conditions "just right":
  - Small changes break either space or time bounds
  - bq-leaves needed for tight space bounds
- Easy to merge together summaries to get summary of union (for distributed computations)
  - Linearity of L and A means everything goes through
- Close to optimal space bounds
  - What about faster updates, less work for queries?
- Made crucial use of tree-structure over universe
  - Can we drop U and work over arbitrary domains?