Algorithms for Primitives of Stream Processing

- Update stream model.

- Stream viewed as sequence (i, v),
  1. i is identity of record.
  2. v is change in frequency of i.
  3. $v > 0$: v insertions of i
  4. $v < 0$: v deletions of i.

Frequency Sketches

- Keys $\in \{0, 1, \ldots, N-1\}$.
- x_i is a random variable for each $i = 0, 1, \ldots, N-1$.
- $\mathbf{P}(x_i = 1) = \mathbf{P}(x_i = -1) = \frac{1}{2}$.
- x_i's 4-wise independent

Frequency Sketches contd.

- Sketch is a random variable X

$$X = \sum_{i=1}^{N-1} f_i x_i$$

- X is efficiently maintained:
  UPDATE(i, v): X **:=** $X + x_i \cdot v$

- $E[X \cdot x_i] = f_i$

- Efficient retrieval of top-k items [Charikar, Chen, Farach 2002]**.**

FM-sketches for Count Distinct Queries

- h is a random hash function from $\{0, \ldots, N-1\} \rightarrow \{0, \ldots, N-1\}$.
- N is a power of 2.
- lsb(x): least significant bit of x.
- $i \rightarrow$ lsb$(h(i))$ ($h(i)$).
- $\mathbf{P}\left(h(i) = 1\right) = \frac{1}{2}, \mathbf{P}\left(h(i) = 2\right) = \frac{1}{4}$
  $\mathbf{P}\left(h(i) = l\right) = \frac{1}{2^l}$

FM-sketches contd.

- Let stream have n distinct items.
- let $l = \lceil \log n \rceil$.
- Expected number of items at level l $= \frac{n}{2^l} \in \{\frac{1}{2}, 1\}$
- Can be used to estimate n.
- Generalizes to all streaming models.