Estimating Statistical Aggregates on Probabilistic Data Streams

T. S. Jayram et al. 2007

James Jolly
March 17, 2010

Probabilistic Data Streams

Events produced by...

- financial markets
- IP networks
- environmental sensors

These are uncertain...

- incomplete knowledge
- stochastic phenomena
- measurement error

Difficulty Computing Aggregates

Want to summarize these events with aggregates...

- min, max, median, etc.
- mean, variance, skew, etc.
- distinct, repeat-rate, etc.

Want to process events online...

- limited working memory
- desire one-pass methods

We are stuck estimating aggregates.

Contribution

Single-pass approximation algorithms for estimating...

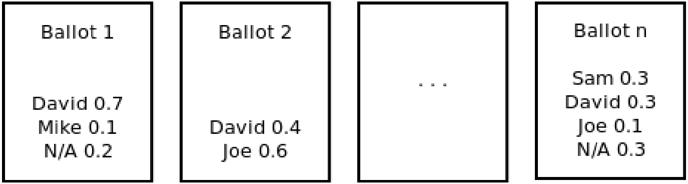
- mean
- median
- distinct
- repeat-rate

... which store a data sketch in memory.

Probabilistic Data Stream Model

Uncertain events...

- are marginal distributions over possible events
- *m* values in domain
- have a probability of not occuring (N/A)
- *n* elements in stream



Ballot 1

David 0.7
Mike 0.1
N/A 0.2

Ballot 2

David 0.4
Joe 0.6

...

Ballot n

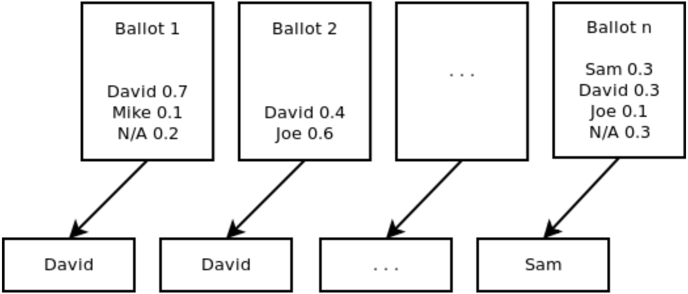
Sam 0.3
David 0.3
Joe 0.1
N/A 0.3

Estimating Distinct (F_0)

- reduced to finding distinct over many deterministic streams
- deterministic streams randomly-generated using marginal probabilties
- distinct value counts approximated in each stream
  Ziv Bar-Yossef et al. 2002

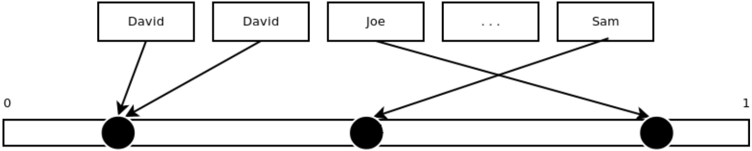
Converting to a Deterministic Stream

- $a_i \in [m]$
- $\forall a_i$, p_i chance of adding each $\left(j, p_i\right) \in a_i$ to new stream



Finding Expected Minimum

- apply random hash function to stream
  $h(j) :\longrightarrow [0,1]$



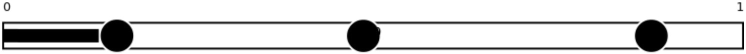
Approximation Intuition

- given expected minimum of n hash values in stream

  $v = \min(h(a_1), h(a_2), ..., h(a_n))$

- if F_0 independent and uniform values in $[0, 1]$

  $$F_0 \overset{\sim}{=} \frac{1}{v}$$



- average F_0 estimates across multiple streams

Approximation Cost

$O\left(log(m)\right)$ in time, $O\left(\frac{1}{\epsilon^2}log(m)\right)$ in space

ϵ - approximation parameter
δ - confidence parameter
$P(F_0^{est} - F_0 \leq \epsilon F_0) \geq 1 - \delta$

Old Techniques Reborn

- desirable to use the results computed in previous queries
- can compute some aggregates from other aggregates
- useful in ‘roll-up’ and ‘drill-down’ operations

Summary

- can estimate aggregates over uncertain data
- cheap to compute in both time and space
- cost is function of domain size