Space-Efficient Estimation of Statistics over Sub-Sampled Streams

Andrew McGregor
University of Massachusetts
mcgregor@cs.umass.edu

A. Pavan
Iowa State University
pavan@cs.iastate.edu

Srikanta Tirthapura
Iowa State University
snt@iastate.edu

David Woodruff
IBM Research Almaden
dpwoodru@us.ibm.com

Sampled IP Packet Streams

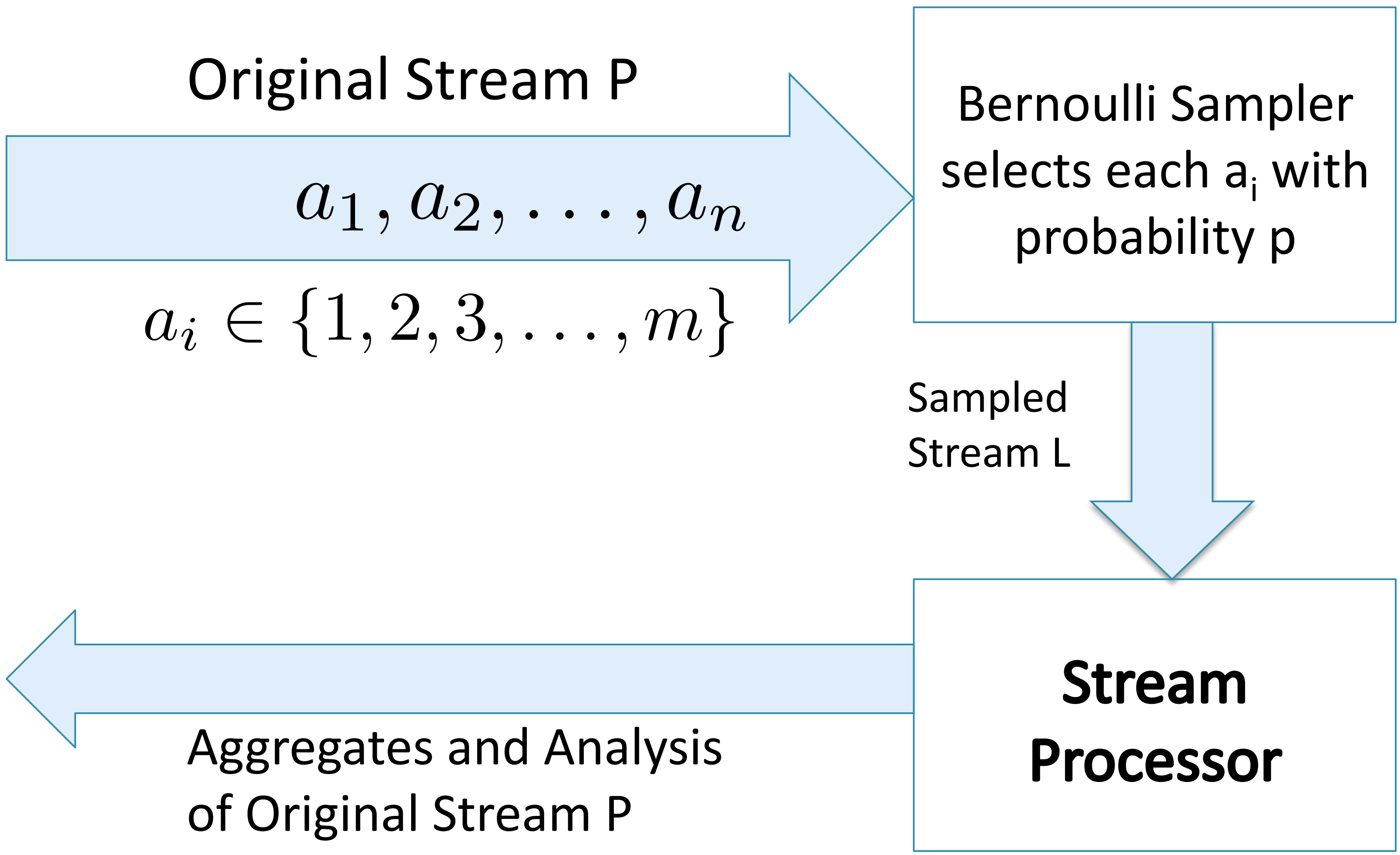
- IP Traffic Monitoring
  - 40Gbps – 100s of Gbps
  - Cisco Netflow standard for monitoring
  - Sampled Netflow
    - Network Monitor sees only a random sample of the original packet stream
    - Different Types of Sampling Used
    - IETF Working Group (psamp)

Sampled Streams

- What can we compute over a stream by observing only a random sample of the stream?

- Two Constraints:
  - Only observe a random sample
  - Streaming, Memory Bound Computation

Model: Bernoulli Sampling



Aggregates

The stream is a sequence of items $(a_1, a_2, \ldots, a_n)$

What matters is the vector $f = \langle f_1, f_2, \ldots, f_m \rangle$ where f_i is frequency of i

- Frequency Moments $F_k(f) = \sum_{i=1}^{m} f_i^k$
- Number of Distinct Elements
- Empirical Entropy $H(f) = \sum_{i=1}^{m} \frac{f_i}{n} \lg\left(\frac{n}{f_i}\right)$
- Heavy Hitters

Preliminaries

- Let g_i be the frequency of item *i* in the substream $g_i = B(f_i, p)$
- *L* contains the frequency vector $\langle g_1, g_2, \ldots, g_m \rangle$
- Randomized multiplicative approximation, for parameters

$$\Pr\left[\frac{1}{\alpha} \leq \frac{X}{\tilde{X}} \leq \alpha\right] \geq 1 - \delta$$

Results

- Number of Distinct Elements
  - (Known) Upper Bound on Result Quality
  - Simple Streaming Algorithm that meets the bound
- Frequency Moments, F_k, $k > 0$
  - Smaller values of *p increase* streaming space complexity
  - Matching upper and lower bounds (w.r.t *p*)
  - Tradeoff between processing time and space
- Entropy
  - Matching Upper and Lower Bounds for Additive Error
  - Relative Error Impossible in small space
- Heavy Hitters
  - Sampling is a good fit

Related Work

- Duffield, Lund, Thorup, “Properties and prediction of flow statistics from sampled packet streams”, IMC 2002
- Duffield, Lund, and Thorup, “Estimating flow distributions from sampled flow statistics”, SIGCOMM 2003
- Rusu and Dobra, “Sketching sampled data streams” ICDE 2009
- Bar-Yossef, “Sampling Lower Bounds via Information Theory”, STOC 2003

Results: Number of Distinct Elements

For constant *p*, estimate $F_0(P)$ from observing L

- Any algorithm must have relative error $\Omega\left(\frac{1}{\sqrt{p}}\right)$ (Charikar et al. PODS 2000)

- A simple streaming algorithm has relative error $O\left(\frac{1}{\sqrt{p}}\right)$

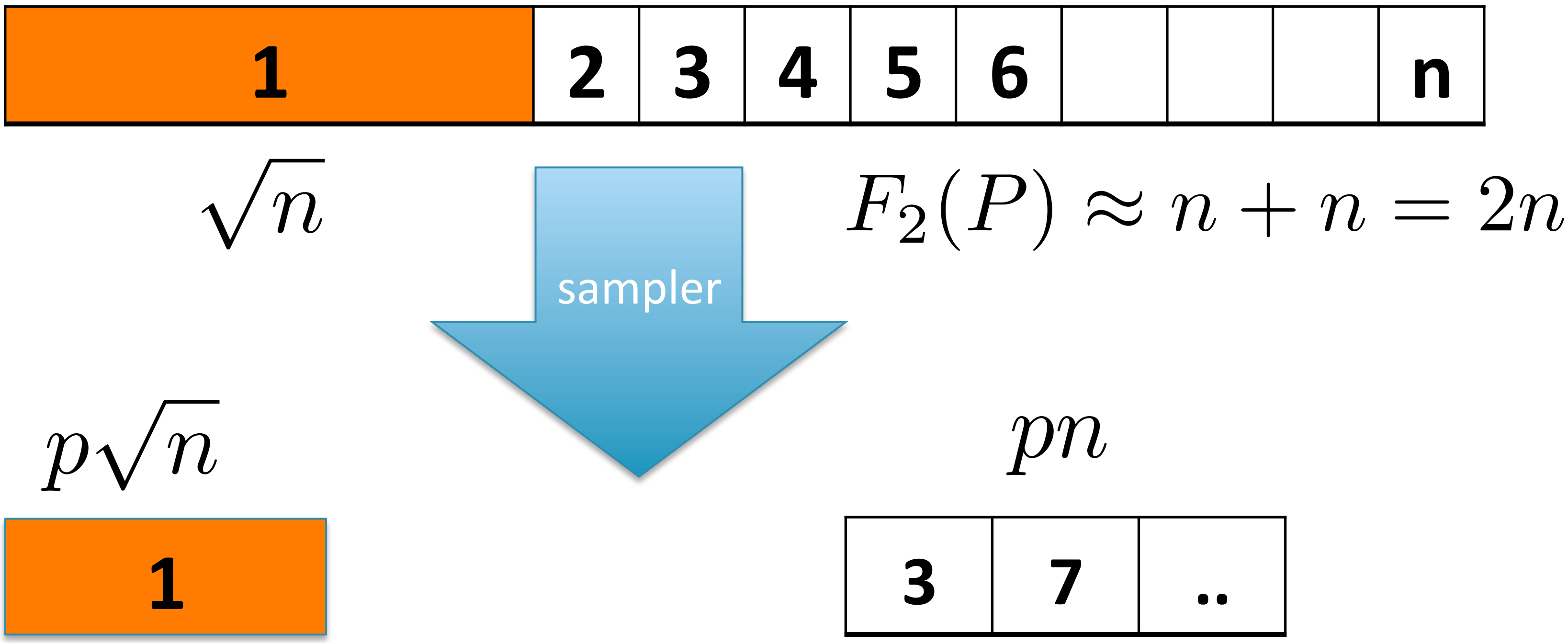
- Other Estimators, such as GEE (Generalized Error Estimator) also possible in single pass

Frequency Moments F_k

Theorem (Upper Bound): There is a one pass streaming algorithm which observes L and outputs a $(1+\epsilon, \delta)$-estimator to $F_k(P)$ where $k \geq 2$ using $\tilde{O}\left(\frac{1}{p} m^{1-2/k}\right)$ space.

assuming $p = \tilde{\Omega}(\min(m,n)^{-1/k})$

Computing F_2: Why is This Hard?



$$F_2(L) = p^2 n + pn = (p^2 + p)n$$

Algorithm 1 for F_2

$$Z = \frac{F_2(L) - F_1(L)}{p^2} \qquad E[Z] = F_2(P)$$

- Algorithm
  1. Estimate $F_2(L)$ using streaming algorithm on L
  2. Use in above formula for Z

Algorithm 1 for F_2 (contd)

- To get $(1+\epsilon, \delta)$-approximation for *Z*, space needed is $\tilde{O}\left(\frac{1}{p^2}\right)$
- Issue: Need to estimate $F_2(L)$ with very high accuracy to get good relative error for

$$\frac{F_2(L) - F_1(L)}{p^2}$$

Algorithm 2 for F_2

- Collisions. The number of 2-wise collisions in P

$$C_2(P) = \sum_{i=1}^{m} \binom{f_i}{2}$$

- Observation: $F_2(P) = 2C_2(P) + F_1(P)$

- Algorithm:
  - Estimate and $C_2(P), F_1(P)$ with mult. error $(1+\epsilon)$

Estimating $C_2(P)$

- Observation: $E[C_2(L)] = p^2 C_2(P)$

  $Var(C_2(L)) = O(p^3 F_2^{1.5})$

- If $C_2(L)$ estimated accurately, we are done

- But this is hard in general, in small space

Estimating $C_2(L)$

$$F_2(P) = 2C_2(P) + F_1(P)$$

Case I:

$$C_2(L) = \Omega(p^2 F_2(P))$$

- Estimate $C_2(L)$ hence $C_2(P)$

with good relative error

Case 2:

$$C_2(L) = O(p^2 F_2(P))$$

$$C_2(P) \ll F_2(P)$$

- Accurate estimate of $C_2(L)$ not needed

- Get an estimate within a multiplicative error of 3

Estimating $C_2(L)$

- Estimate with good relative error when

$$C_2(L) = \Omega(p^2 F_2(P))$$

- Technique due to Indyk & Woodruff (STOC 2005)
  - Divide items *{1,2..,m}* into classes based on frequency
  - Estimate the size of different classes that contribute to the final result

Tight Lower Bound for F_k

Theorem: Any constant-pass streaming algo. that $(1+\epsilon,\delta)$-approximates F_k for a sufficiently small constants ϵ,δ by observing a sampled stream, in the Bernoulli sampling model, requires $\Omega\left(m^{1-2/k}/p\right)$ bits of space

F_k Time-Space Tradeoff

With sampled stream at probability *p*

- Processing Time: $\tilde{O}(pn)$
- Streaming Space: $\tilde{O}\left(\frac{m^{1-2/k}}{p}\right)$
- Product: $\tilde{O}(nm^{1-2/k})$

Entropy

- No multiplicative error approximation possible with probability 9/10, even if p > ½
  - The entropy of sampled stream could be zero, while that of the original stream non-zero
- If $p = \Omega(n^{-1/3})$ there is an approximation H' to $H(f)$ such that
  - $H' \leq 100H(f)$ with prob. at least 99/100

    $H' \geq H(f)/2 - o(1)$

Heavy Hitters

- The frequency of heavy hitters are (approximately) proportionately maintained in the sampled stream
- Precise upper and lower bounds in paper

Conclusions

- Extreme Volume Data Streams
- F_k
  - Upper and Lower Bounds for F_k, $k > 0$
  - Smooth Space-Time tradeoff
- Number of distinct Elements, Entropy
  - Sampling harms, streaming does not
- Heavy Hitters:
  - Sampling and Streaming are both ok