SKETCHING SAMPLED DATA STREAMS

FLORIN RUSU and **ALIN DOBRA**
CISE Department, University of Florida, Gainesville, FL, USA

Analyze **fast** data streams
→ single pass
→ fixed order
→ small space

Aggregates over joins
→ *Skew* of a relation read from the disk
→ *Correlation* between flows passing through a high-speed router

SAMPLING

Stream F: a | 1 1 2 3 1 3, frequency vector $\mathbf{f}$:

i	1	2	3
f_i	3	1	2

Sample F': a | 1 3 1, sampled frequency vector $\mathbf{f}'$:

i	1	2	3
f'_i	2	0	1

Stream G: a | 3 1 3 1 1, frequency vector $\mathbf{g}$:

i	1	2	3
g_i	3	0	2

Sample G': a | 3 1, sampled frequency vector $\mathbf{g}'$:

i	1	2	3
g'_i	1	0	1

$$X = C \cdot \mathbf{f}' \cdot \mathbf{g'^T} = C \cdot \begin{bmatrix}2 & 0 & 1\end{bmatrix} \cdot \begin{bmatrix}1\\0\\1\end{bmatrix} = C \cdot 3 \approx 13$$

Random sampled frequency vector (**moment generating function**)

$$E[X] = C \sum_{i \in I} E[f'_i] E[g'_i]$$

$$\mathrm{Var}[X] = C^2 \left[\sum_{i \in I} \sum_{j \in I} E\left[f'_i f'_j\right] E\left[g'_i g'_j\right] - \left(\sum_{i \in I} E[f'_i] E[g'_i] \right)^2 \right]$$

SKETCHES

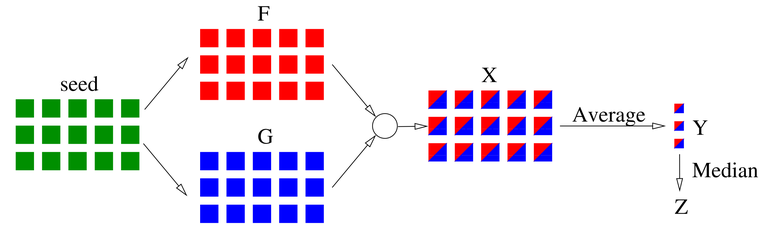
$$\xi = \begin{bmatrix}\xi_1 & \xi_2 & \xi_3\end{bmatrix} = \begin{bmatrix}-1 & +1 & -1\end{bmatrix}$$

$$X_F = \mathbf{f}\xi^T = \begin{bmatrix}3 & 1 & 2\end{bmatrix} \begin{bmatrix}-1\\+1\\-1\end{bmatrix} = -4, X_G = \mathbf{g}\xi^T = \begin{bmatrix}3 & 0 & 2\end{bmatrix} \begin{bmatrix}-1\\+1\\-1\end{bmatrix} = -5$$

$$X = X_F X_G = (-4)(-5) = 20 \approx 13$$

ξ is a family of 4-wise independent random variables

$$E[X] = E[\mathbf{f}\xi^T \xi \mathbf{g}^T] = \mathbf{f} E[\xi^T \xi] \mathbf{g}^T = \mathbf{f} I \mathbf{g}^T = \mathbf{f}\mathbf{g}^T, \mathrm{Var}[X] = \sum_{i \in I} f_i^2 \sum_{j \in I} g_j^2 + \left(\sum_{i \in I} f_i g_i \right)^2 - 2 \sum_{i \in I} f_i^2 g_i^2$$



SKETCHES OVER SAMPLES

$$X'_F = \mathbf{f}'\xi^T = \begin{bmatrix}2 & 0 & 1\end{bmatrix} \begin{bmatrix}-1\\+1\\-1\end{bmatrix} = -3, X'_G = \mathbf{g}'\xi^T = \begin{bmatrix}1 & 0 & 1\end{bmatrix} \begin{bmatrix}-1\\+1\\-1\end{bmatrix} = -2$$

$$X = C \cdot X'_F X'_G = C \cdot (-3)(-2) = C \cdot 6 \approx 13$$

Build the sketch over non-materialized samples

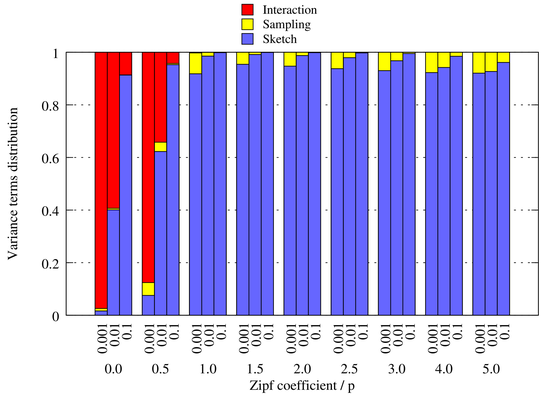
$$X = C \cdot \sum_{i \in I} f'_i \xi_i \cdot \sum_{j \in I} g'_j \xi_j$$

$$E[X] = C \cdot \sum_{i \in I} E[f'_i] E[g'_i]$$

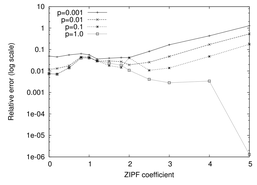
$$\mathrm{Var}[X] = C^2 \cdot \left[\sum_{i \in I} E\left[f'^2_i\right] \sum_{j \in I} E\left[g'^2_j\right] + 2 \cdot \sum_{i \in I} \sum_{j \in I} E\left[f'_i f'_j\right] E\left[g'_i g'_j\right] - 2 \cdot \sum_{i \in I} E\left[f'^2_i\right] E\left[g'^2_i\right] - \left(\sum_{i \in I} E[f'_i] E[g'_i] \right)^2 \right]$$

$$\mathrm{Var}\left[\frac{1}{n} \cdot \sum_{k=1}^{n} X_k\right] = \frac{1}{n}\left[\mathrm{Var}[X_k] + (n-1) \cdot \mathrm{Cov}_{k \neq l}[X_k, X_l]\right]$$

$$\mathrm{Var}_{\text{sketch over samples}} = \mathrm{Var}_{\text{sketch}} + \mathrm{Var}_{\text{sampling}} + \mathrm{Var}_{\text{interaction}}$$



Bernoulli sampling



Sampling without replacement