**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
- $\overline{\text{Skew of a relation read from the disk}}$
- Correlation between flows passing through a high-speed router

**SAMPLING**

Stream $F$: $\{1 2 3 1 3\}$ frequency vector $f$: $\{1 2 3\}$

Sample $F'$: $\{1 3\}$ sampled frequency vector $f'$: $\{1 2 3\}$

Stream $G$: $\{1 2 3 1 3\}$ frequency vector $g$: $\{1 2 3\}$

Sample $G'$: $\{1 3\}$ sampled frequency vector $g'$: $\{1 2 3\}$

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

Random sampled frequency vector (moment generating function)

$E[X] = C \sum_{i=1}^{n} E[f_i]E[g_i]$

$\text{Var}[X] = C^3 \left[ \sum_{i \neq j} E[f_i]E[g_j] - \left( \sum_{i=1}^{n} E[f_i]E[g_i] \right)^2 \right]$

$X_F = f'^T = \begin{bmatrix} 2 & 0 & 1 \end{bmatrix}$

$X_G = g'^T = \begin{bmatrix} 1 & 3 \end{bmatrix}$

$X = X_F X_G = \begin{bmatrix} -4 \end{bmatrix} \cdot \begin{bmatrix} -0 \end{bmatrix} = 20 \approx 13$

$\xi = [\xi_1, \xi_2, \xi_3, \xi_4] = [-1, 1, -1, -1]$

$X_F = f'^T = \begin{bmatrix} 3 & 1 \end{bmatrix}$

$X_G = g'^T = \begin{bmatrix} 0 & 2 \end{bmatrix}$

$X = X_F X_G = \begin{bmatrix} -4 \end{bmatrix} \cdot \begin{bmatrix} -0 \end{bmatrix} = 20 \approx 13$

$\xi$ is a family of 4-wise independent random variables

$E[X] = E[f'^Tg'^T] = f'^Tg'^T = f'^Tg'^T = f'^Tg'^T, Var[X] = \sum_{i \neq j} E[f_i]E[g_j] + \left( \sum_{i=1}^{n} f_i g_i \right)^2 - 2 \sum_{i \neq j} f_i g_j$

**SKETCHES**

**SKETCHES OVER SAMPLES**

$\text{Var}\left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n} \text{Var}[X] + (n - 1) \cdot \text{Cov}_{2,2}[X_i, X_i]$

$\text{Var}\text{Sketch over samples} = \text{VarSketch} + \text{VarSample} + \text{VarInteraction}$

Bernoulli sampling

Sampling without replacement

---

**REFERENCES**

- Analyze fast data streams
- Aggregates over joins
- Sampling from large datasets
- Correlation between flows passing through a high-speed router

---

**ACKNOWLEDGMENTS**

- Funding from the National Science Foundation
- Support from the University of Florida

---

**CONCLUSIONS**

- Analyze fast data streams
- Aggregates over joins
- Sampling from large datasets