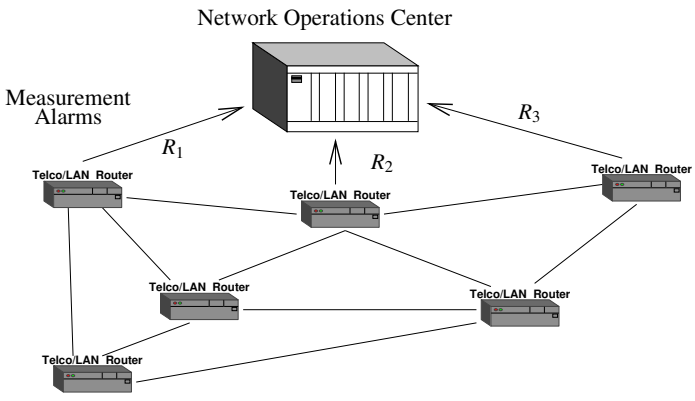
Fast Range-Summable Random Variables for Efficient Aggregate Estimation

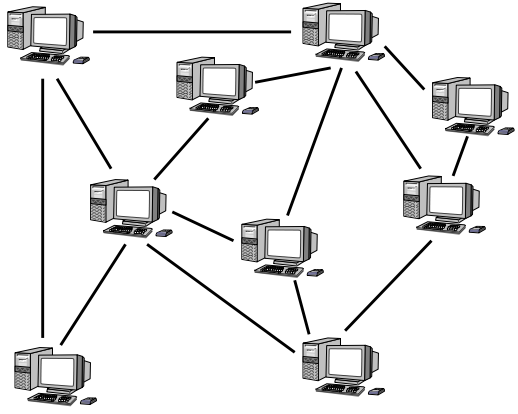
Florin Rusu and Alin Dobra
CISE Department
University of Florida

June 27, 2006

Data-Streaming and Distributed Computation



Network Monitoring



Peer-to-peer/sensor network

- Large amount of data at high speeds
- Sketches summarize data in small space
- MEMORY

- Communication is energy consuming
- Sketches are linear $\rightarrow$ Send sketches
- COMMUNICATION

Outline of the Talk

- AMS Sketches
  - 4-Wise Independent Random Variables
- Motivating Applications
- Existing Solutions
- Proposed Solution
- Experimental Evaluation

AMS Sketches (AMS 1996, AGMS 1999)

Problem

- Estimate the size of join of two relations F and G, $|F \bowtie_a G|$, in small space
- Solution: Summarize the data in each relation as a single number X_F, X_G

Sketches

- Select random variables $\xi_1, \xi_2, \xi_3 \in \{-1, +1\}$
- Relation F:

a	1	1	2	3	1	3

$$X_F = \xi_1 + \xi_1 + \xi_2 + \xi_3 + \xi_1 + \xi_3$$

- Relation G:

a	3	1	3	1	1

$$X_G = \xi_3 + \xi_1 + \xi_3 + \xi_1 + \xi_1$$

- $X = X_F X_G$ estimates $|F \bowtie_a G|$

Analysis of AMS Sketches (1)

Random Vectors

- $\xi = [\xi_1 \ldots \xi_n]$, $\xi_i \in \{-1, +1\}$ with $E[\xi_i] = 0$

Sketches

- $X_F = \sum_{t \in F} \xi_{t.a} = \sum_i f_i \xi_i$
- $X_G = \sum_{t \in G} \xi_{t.a} = \sum_i g_i \xi_i$

Size of Join Estimator

- $X = X_F X_G$ is unbiased if ξ is 2-wise independent, i.e., $E[\xi_i \xi_j] = E[\xi_i] E[\xi_j] = 0$

$$E[X] = E\left[\left(\sum_i \xi_i f_i\right)\left(\sum_i \xi_i g_i\right)\right] = E\left[\sum_i {\xi_i}^2 f_i g_i + \sum_{i \neq j} \xi_i \xi_j f_i g_j\right]$$

$$E[X] = \sum_i f_i g_i = |F \bowtie_a G|$$

Analysis of AMS Sketches (2)

Estimator Variance

$$\text{Var}[X] = E\left[X^2\right] - E[X]^2$$

$$E\left[X^2\right] = E\left[\left(\sum_i f_i\xi_i \sum_i g_i\xi_i\right)^2\right] = \sum_i\sum_j\sum_k\sum_l f_i f_j g_k g_l E\left[\xi_i\xi_j\xi_k\xi_l\right]$$

$$E\left[X^2\right] = \sum_i\sum_j f_i^2 g_j^2 + 2\cdot\left(\sum_i f_i g_i\right)^2 - 2\cdot\sum_i f_i^2 g_i^2 + \sum_i\sum_j\sum_k\sum_{i\neq j\neq k\neq l} f_i f_j g_k g_l E\left[\xi_i\xi_j\xi_k\xi_l\right]$$

- If ξ is 4-wise independent, i.e., $E\left[\xi_i\xi_j\xi_k\xi_l\right] = E[\xi_i]E\left[\xi_j\right]E[\xi_k]E[\xi_l] = 0$

$$\text{Var}[X] \leq 2\left(\sum_i {f_i}^2 \sum_j {g_j}^2 - \sum_i {f_i}^2 {g_i}^2\right)$$

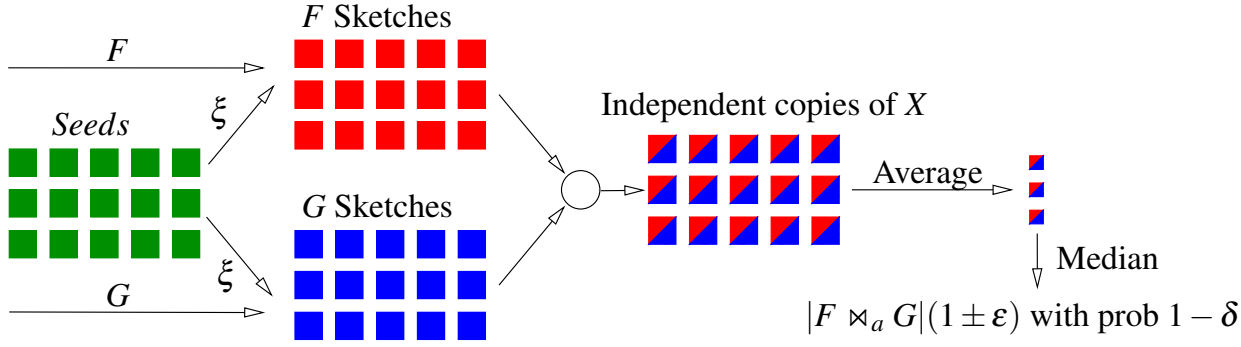
$$\text{Var}[X] \leq 2\ \text{SJ}(F)\ \text{SJ}(G)$$

Sketch Error Reduction

- $|F \bowtie_a G|$ estimation from single sketches of F and G is *noisy*

Solution

- Average $\frac{8}{\varepsilon^2}\frac{\mathrm{Var}[X]}{E[X]^2}$ independent copies of X to reduce the error to ε
- Compute median of $2\log\frac{1}{\delta}$ such averages to increase the confidence to $1-\delta$



4-Wise Independent Random Variables

Problem

- Domain $I = \{0, 1, \ldots, 2^n - 1\},\ n \geq 0$
- Family of ± 1 random variables $\xi\{I\}$
  - $\xi_i(S) = (-1)^{f(s,i)}$
  - S - seed domain, $f(s,i)$ - generating function

Definition

- Uniform 4-wise independent family ξ [CGLMS 2005]
  - $\forall i_1, i_2, i_3, i_4 \in I,\ i_1 \neq i_2 \neq i_3 \neq i_4$
  - $\forall v_1, v_2, v_3, v_4 \in \{-1, +1\}$

$$Pr[\xi_{i_1}(S) = v_1 \wedge \xi_{i_2}(S) = v_2 \wedge \xi_{i_3}(S) = v_3 \wedge \xi_{i_4}(S) = v_4] = \frac{1}{2^4} = \frac{1}{16}$$

Generating Schemes

BCH Codes [ABI 1986, HSS 1999]

- Dense - seed size is $2n+1$
- Exponentiation over a finite field

Reed-Muller Codes [HSS 1999, CGLMS 2005]

- Seed size is large, e.g., $1+n+\frac{n(n-1)}{2}$
- Fast range-summable

Polynomials over Primes [CW 1979]

- Computations over a prime field

Tabulation Based [TZ 2003]

- Cache size constrained

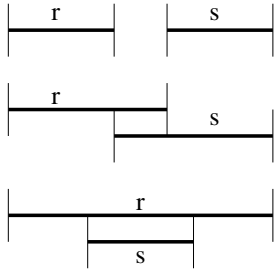
Outline of the Talk

- AMS Sketches
- Motivating Applications
  - Generic Problem
- Existing Solutions
- Proposed Solution
- Experimental Evaluation

Size of Spatial Join

Problem

- R, S sets of intervals: $[l(r_i), u(r_i)]$, $[l(s_j), u(s_j)]$
- $|R \bowtie_o S| = |\{(r,s) | r \in R \wedge s \in S \wedge \text{overlap}(r,s)\}|$



Solution [DGR 2004]

- Relations
  - Interval: R_I, S_I
  - Endpoint: R_E, S_E
- Size of Join
  - $|R \bowtie_o S| = \frac{1}{2}[|R_E \bowtie S_I| + |R_I \bowtie S_E|]$

Selectivity Estimation for Dynamic Histograms

Problem

- Streaming relation $R(A)$, $A \in \{1 \ldots n\}$
- Histogram H for the frequency distribution $D : \{1 \ldots n\} \rightarrow \{1 \ldots M\}$
  - $\{(S_1, v_1) \ldots (S_k, v_k)\}$, $S_i = [\alpha_i, \beta_i] \subset \{1 \ldots n\}$
  - $\min_H ||D - H||_2$

Solution [TGIK 2002]

- Relations
  - $R(S_i) = \begin{cases} 1 & \text{if} \quad j \in S_i \\ 0 & \text{otherwise} \end{cases}$
- Size of Join
  - $v_i = |R \ltimes R(S_i)|$

Class of Applications

Spatial Size of Join, Selectivity Estimation

- Reduction to Size of Join
- Solution based on AMS Sketches
- Extra Requirement: $\sum_{i\in[\alpha,\beta]} \xi_i$
  - Interval input updates
  - Interval queries

Solutions to $\sum_{i\in[\alpha,\beta]} \xi_i$

- Generate each ξ_i and sum
- Fast Range-Summing [GKMS 2003, CGLMS 2005]
- Dyadic Mapping DMAP [GKMS 2002, CM 2004, DGR 2004]

Generic Problem

Interval Sketches

- Select random variables $\xi_1, \xi_2, \xi_3 \in \{-1, +1\}$
- Relation F:

a	[1, 3]	[1, 2]	[2, 3]	[3, 3]	[1, 2]

$$X_F = \sum_{i=1}^{3} \xi_i + \sum_{i=1}^{2} \xi_i + \sum_{i=2}^{3} \xi_i + \sum_{i=3}^{3} \xi_i + \sum_{i=1}^{2} \xi_i$$

- Relation G:

a	3	1	3	1	1

$$X_G = \xi_3 + \xi_1 + \xi_3 + \xi_1 + \xi_1$$

- $X = X_F X_G$ estimates $|F \bowtie_a G|$

Outline of the Talk

- AMS Sketches
- Motivating Applications
- Existing Solutions
  - Dyadic Mapping DMAP
- Proposed Solution
- Experimental Evaluation

4-Wise Independent Fast Range-Summable Random Variables

Problem

- $I = \{0, 1, \ldots, 2^n - 1\},\ n \geq 0$
- $\xi\{I\},\ \xi_i(S) = (-1)^{f(s,i)}$
- Interval $[\alpha, \beta],\ \alpha, \beta \in I,\ \alpha \leq \beta$

$$g([\alpha, \beta], s) = \sum_{\alpha \leq i \leq \beta} \xi_i(s) = \sum_{\alpha \leq i \leq \beta} (-1)^{f(s,i)},\ s \in S$$

Definition

- Bitwise fast range-summable family [CGLMS 2005]
  - $g([\alpha, \beta], s)$ is $O(\log^{O(1)}(\beta - \alpha))$

Generating Schemes

Method [L 2005]

- Represent $f(s,i)$ as an XOR-AND boolean formula
- Results in assignment satisfiability counting [EK 1990]
  - 3XOR-AND formulae are #P-complete
  - Algorithm for 2XOR-AND formulae

Not Fast Range-Summable Schemes

- BCH Codes, Polynomials over Primes
- Tabulation Based

Fast Range-Summable Scheme

- Reed-Muller Codes [GKMS 2003, CGLMS 2005]
  - $O(\log^4 n)$, 40 interval sketches/second
  - Impractical

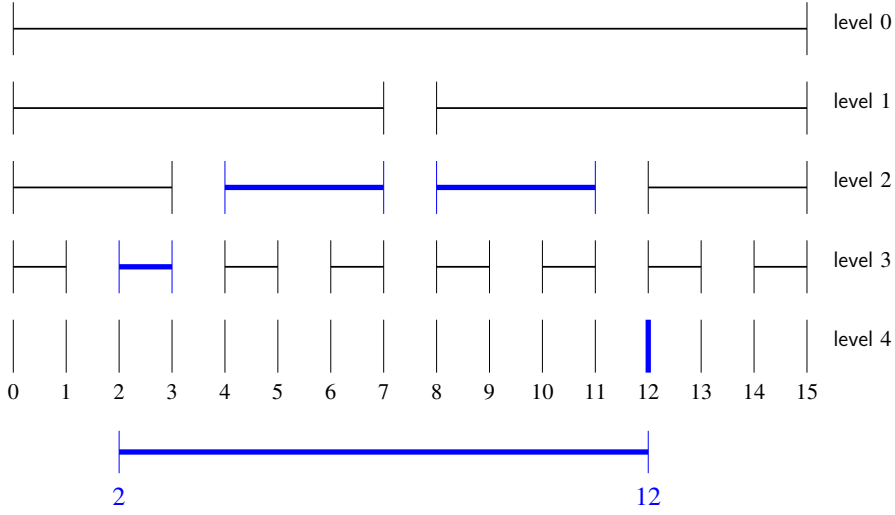
Dyadic Mapping DMAP (1)

Dyadic Intervals [GKMS 2002]

- $[q2^j, (q+1)2^j)$, $0 \le j \le n$, $0 \le q \le 2^{n-j} - 1$

Minimal Dyadic Cover

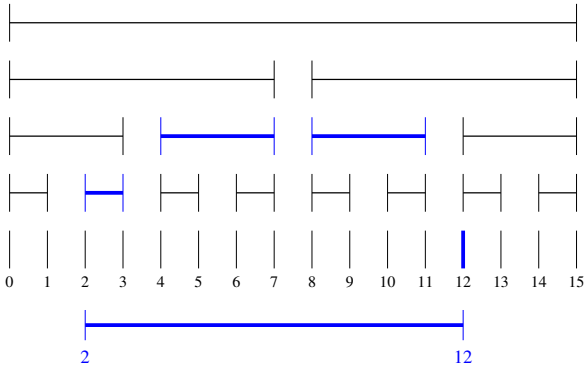
- $D([\alpha, \beta)) = [\alpha = \alpha_1, \beta_1) \cup [\alpha_2, \beta_2) \cup \cdots \cup [\alpha_k, \beta_k = \beta)$
- $|D([\alpha, \beta))|$ is $O(\log(\beta - \alpha))$



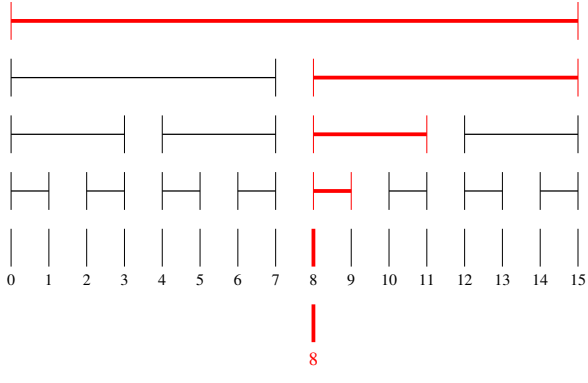
Dyadic Mapping DMAP (2)

Mappings [DGR 2004]

- Domain $I \rightarrow$ Dyadic intervals over I
- Interval $[\alpha, \beta) \rightarrow$ Minimal dyadic cover $D([\alpha, \beta))$
- Point $\gamma \rightarrow$ Dyadic intervals containing γ



F_M



G_M

Dyadic Mapping DMAP (3)

- $|F \bowtie G| = |F_M \bowtie G_M|$

Analysis

- $\mathrm{Var}\,[|F \bowtie G|] \leq 2\ \mathrm{SJ}(F)\ \mathrm{SJ}(G) = 2\sum_i f_i^2 \sum_i g_i^2$
- $\mathrm{Var}\,[|F_M \bowtie G_M|] \leq 2\ \mathrm{SJ}(F_M)\ \mathrm{SJ}(G_M) = 2\sum_\delta f_\delta^2 \sum_\delta g_\delta^2$

$[\alpha,\beta)$	$(\beta-\alpha)f_i$	$O(\log(\beta-\alpha))f_\delta$
γ	g_i	$O(\log n)g_\delta$

- $\sum_i f_i^2 \geq \sum_\delta f_\delta^2$, $\sum_i g_i^2 \leq \sum_\delta g_\delta^2$
- $\mathrm{Var}\,[|F \bowtie G|]$ not comparable $\mathrm{Var}\,[|F_M \bowtie G_M|]$

Outline of the Talk

- AMS Sketches
- Motivating Applications
- Existing Solutions
- Proposed Solution
  - Extended Hamming Scheme EH3
- Experimental Evaluation

AMS Sketches Variance

$$\text{Var}[X] = E\left[X^2\right] - E[X]^2$$

$$E\left[X^2\right] = E\left[\left(\sum_i f_i\xi_i \sum_i g_i\xi_i\right)^2\right] = \sum_i\sum_j\sum_k\sum_l f_i f_j g_k g_l E\left[\xi_i\xi_j\xi_k\xi_l\right]$$

$$E\left[X^2\right] = \sum_i\sum_j f_i^2 g_j^2 + 2\cdot\left(\sum_i f_i g_i\right)^2 - 2\cdot\sum_i f_i^2 g_i^2 + \sum_i\sum_j\sum_k\sum_{i\neq j\neq k\neq l} f_i f_j g_k g_l E\left[\xi_i\xi_j\xi_k\xi_l\right]$$

- $E[\xi_i] = 0$
- If ξ is $4-$wise independent, i.e., $E\left[\xi_i\xi_j\xi_k\xi_l\right] = E[\xi_i]E\left[\xi_j\right]E[\xi_k]E[\xi_l] = 0$

$$\text{Var}[X] = \sum_i\sum_j f_i^2 g_j^2 + \left(\sum_i f_i g_i\right)^2 - 2\cdot\sum_i f_i^2 g_i^2$$

BCH3 Scheme (ABI 1986)

- $\xi_i(S) = (-1)^{f(s,i)},\ f(s,i) = s_0 \oplus S_1 \cdot i$

$$
\begin{aligned}
E\left[\xi_i\xi_j\xi_k\xi_l\right] &= E\left[(-1)^{s_0\oplus S_1\cdot i}\cdot(-1)^{s_0\oplus S_1\cdot j}\cdot(-1)^{s_0\oplus S_1\cdot k}\cdot(-1)^{s_0\oplus S_1\cdot l}\right] \\
&= E\left[(-1)^{S_1\cdot i\oplus S_1\cdot j\oplus S_1\cdot k\oplus S_1\cdot l}\right] \\
&= E\left[(-1)^{S_1\cdot(i\oplus j\oplus k\oplus l)}\right]
\end{aligned}
$$

$$
E\left[\xi_i\xi_j\xi_k\xi_l\right] = E\left[(-1)^{S_1\cdot(i\oplus j\oplus k\oplus l)}\right] = \begin{cases} 0 & \text{if } i\oplus j\oplus k\oplus l \neq \bar{0} \\ 1 & \text{if } i\oplus j\oplus k\oplus l = \bar{0} \end{cases}
$$

$$
\Delta\mathrm{Var}\left[\mathrm{BCH3}\right] = \sum_i \sum_j \sum_{k\neq i\neq j} f_i f_j g_k g_{i\oplus j\oplus k}
$$

EH3 Scheme (FKSV 2002)

- $\xi_i(S) = (-1)^{f(s,i)}$, $f(s,i) = s_0 \oplus S_1 \cdot i \oplus h(i)$

$$E\left[\xi_i\xi_j\xi_k\xi_l\right] = E\left[(-1)^{S_1\cdot(i\oplus j\oplus k\oplus l)\oplus(h(i)\oplus h(j)\oplus h(k)\oplus h(l))}\right]$$
$$= \begin{cases} 0 & \text{if } i\oplus j\oplus k\oplus l \neq \bar{0} \\ 1 & \text{if } (i\oplus j\oplus k\oplus l = \bar{0}) \wedge (h(i)\oplus h(j)\oplus h(k)\oplus h(l) = 0) \\ -1 & \text{if } (i\oplus j\oplus k\oplus l = \bar{0}) \wedge (h(i)\oplus h(j)\oplus h(k)\oplus h(l) = 1) \end{cases}$$

- $h(i) = (i_0 \vee i_1) \oplus (i_2 \vee i_3) \oplus \cdots \oplus (i_{n-2} \vee i_{n-1})$
- Average case analysis for f_i not correlated with g_j

$$E\left[\Delta\text{Var}\left[\text{EH3}\right]\right] = Q\frac{1}{|I|}\left(\sum_i f_i\right)^2\left(\sum_i g_i\right)^2$$

- Best case analysis for $f_i = f$, $g_j = g$, $|I| = 4^n$

$$\text{Var}\left[\text{EH3}\right] = 0$$

EH3 is Fast Range-Summable

Algorithm $O(\log(\beta - \alpha))$

- Compute the minimal dyadic cover $D([\alpha,\beta)) = \{[\alpha_1,\beta_1) \cup \cdots \cup [\alpha_d,\beta_d)\}$
- For each interval in $D([\alpha,\beta))$ compute $g_j = g([\alpha_j,\beta_j),s) = \sum_{\alpha_j \le i < \beta_j} \xi_i(s)$
- $g([\alpha,\beta),s) = \sum_{j=1}^{d} g_j$

Sketching Intervals

- Reed-Muller: 40 interval sketches/second
- DMAP: $600,000$ interval sketches/second
- EH3: $500,000$ interval sketches/second

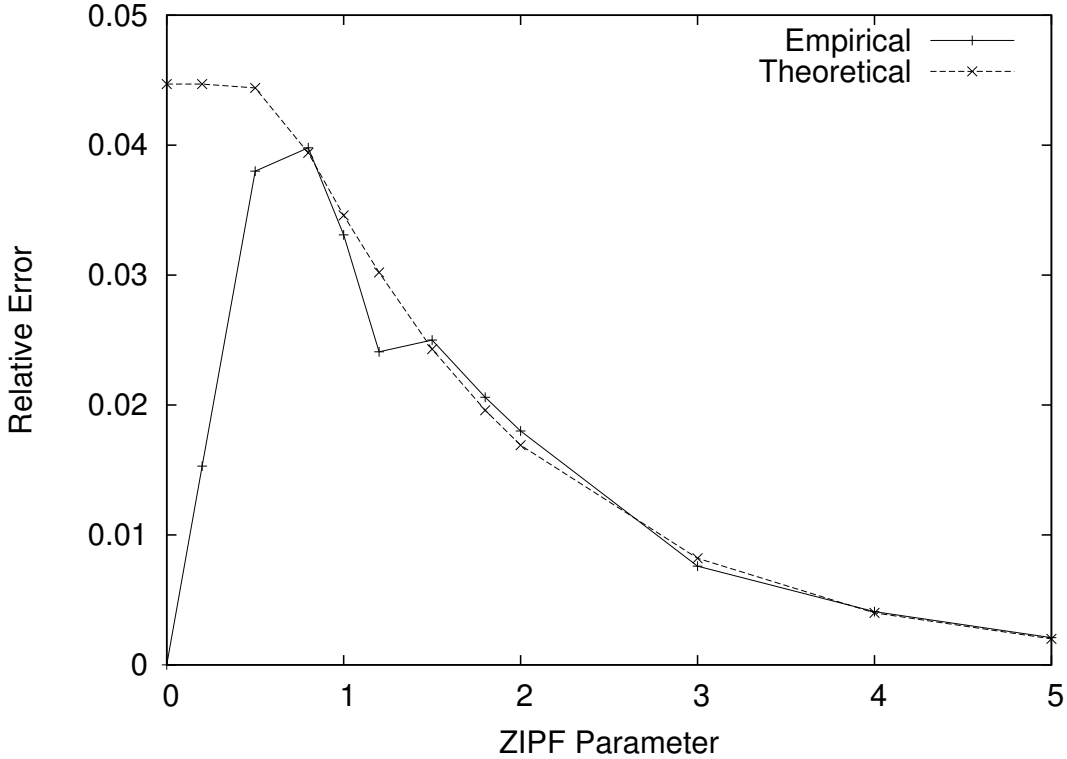
Outline of the Talk

- AMS Sketches
- Motivating Applications
- Existing Solutions
- Proposed Solution
- Experimental Evaluation

EH3 Variance Validation

Settings

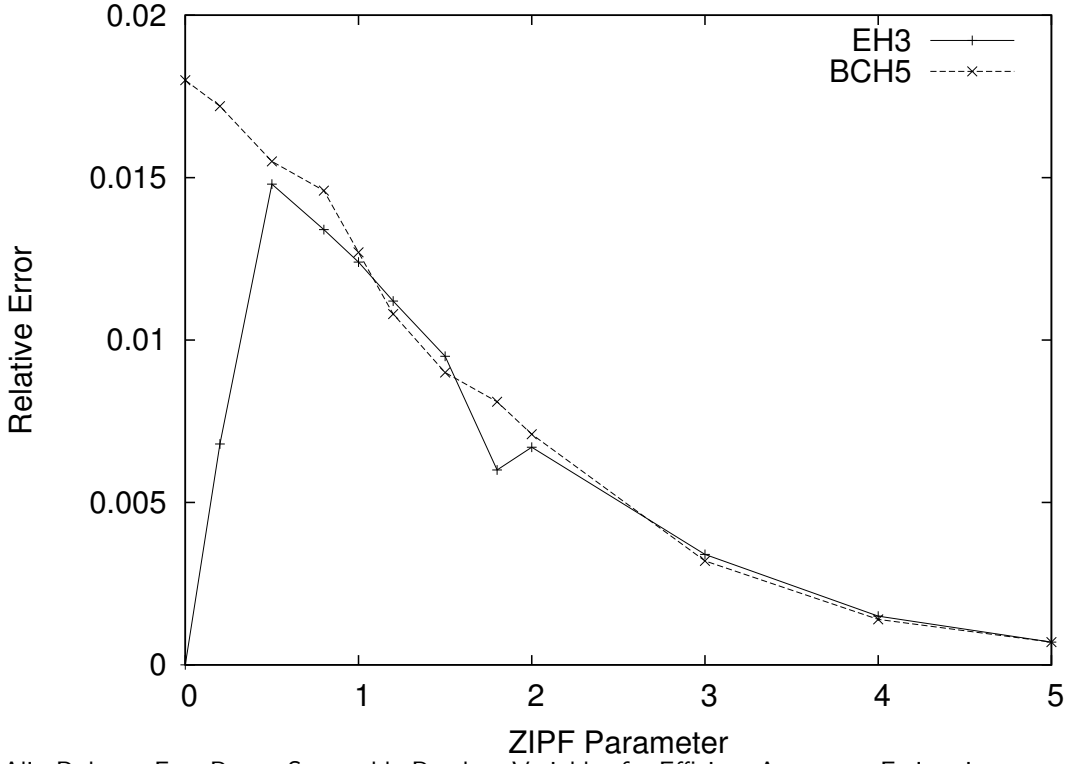
- $|I| = 2^{14} = 16,384$, $100,000$ tuples, $\frac{|\text{Estimated SJ}(F) - \text{SJ}(F)|}{\text{SJ}(F)}$



EH3 vs BCH5

Settings

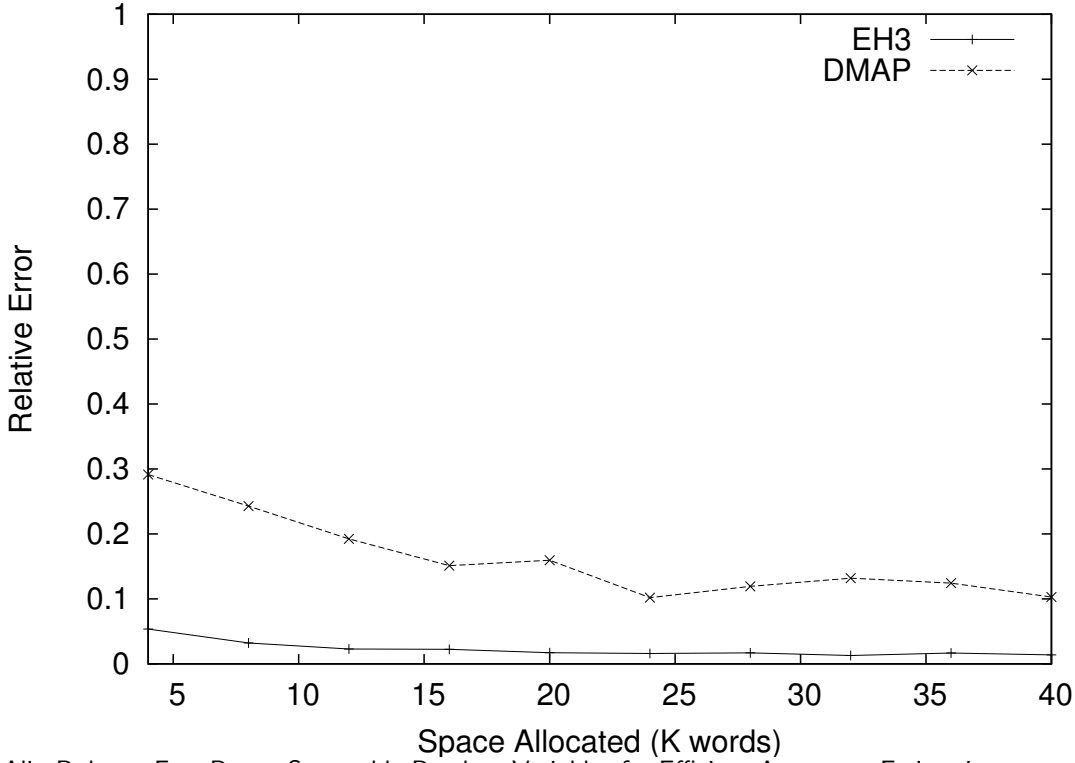
- $|I| = 2^{14} = 16,384$, $100,000$ tuples, 10 medians, $\frac{|\text{Estimated SJ}(F) - \text{SJ}(F)|}{\text{SJ}(F)}$



EH3 vs DMAP for Size of Spatial Join

Settings

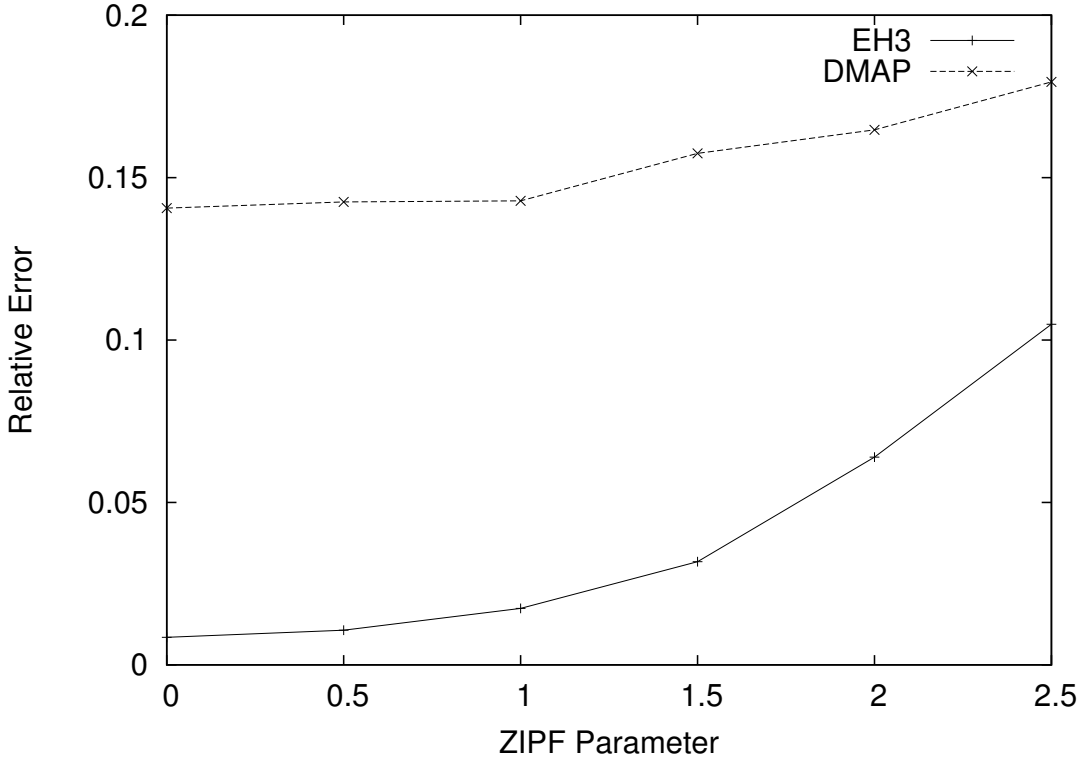
- LANDC $14,731$ tuples, SOIL $29,662$ tuples, $\frac{|\text{Estimated } |\text{LANDC}\bowtie\text{SOIL}| - |\text{LANDC}\bowtie\text{SOIL}||}{|\text{LANDC}\bowtie\text{SOIL}|}$



EH3 vs DMAP for Selectivity Estimation

Settings

- 1024×1024, 10 regions, $\frac{|\text{Estimated Selectivity} - \text{Selectivity}|}{\text{Selectivity}}$



Conclusions

Contributions

- Analysis of the ± 1 random variables generating schemes
- Identify the size of join between intervals and points as a generic problem
- Analysis of the fast range-summation property for the 4-wise schemes
- Identify DMAP as a generic method
- Detailed analysis of the AMS sketches variance
- Extensive experimental study

EH3 Scheme for Size of Join Estimations using AMS-Sketches

- Simple to compute
- Small variance
- Fast range-summable

Questions