

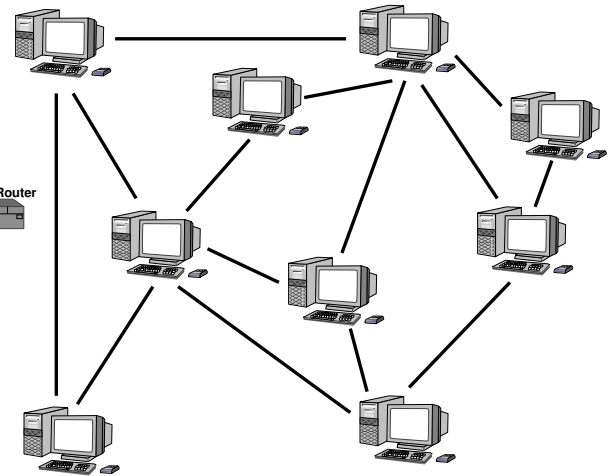
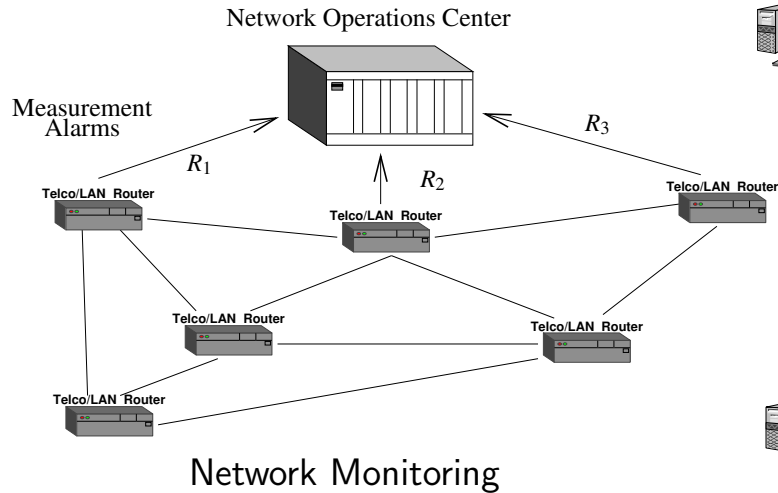
Sketches for Aggregate Estimations over Data Streams

Florin Rusu

Advisor: Alin Dobra

February 4, 2008

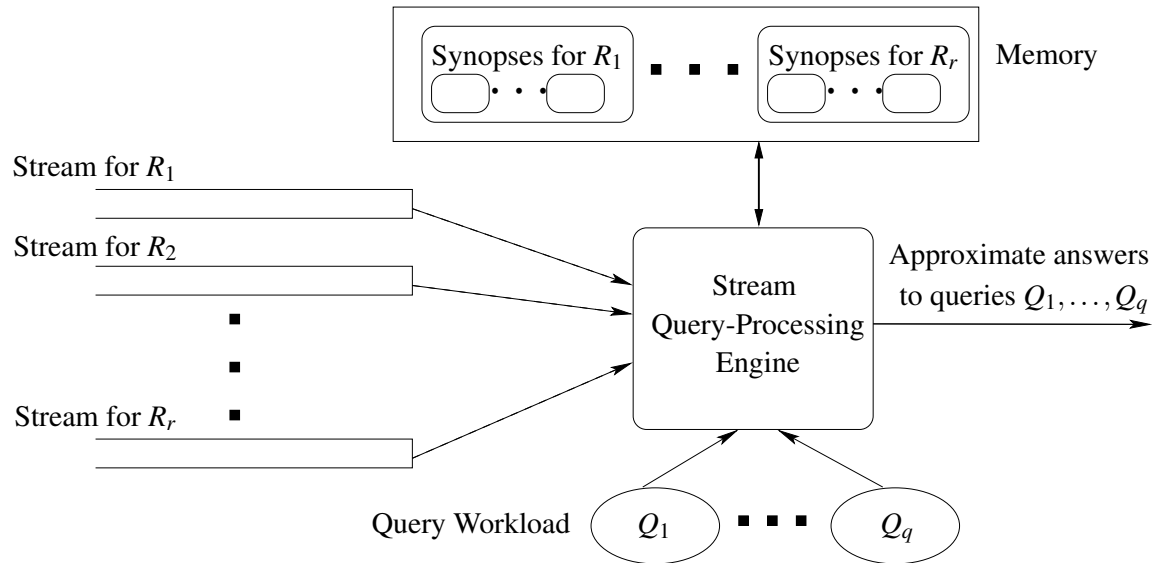
Data Streaming and Distributed Computation



Distributed Networks

- Large amount of data at high speeds
- Sketches summarize data in small space
- MEMORY
- Communication is energy consuming
- Sketches are linear → Send sketches
- COMMUNICATION

Streaming Computational Model



• Computational model

- *Single Pass*: each record examined at most once
- *Fixed Order*: no assumption about the arrival order can be made
- *Small Space*: unbounded size stream
- *Real-time*: per-record processing time must be low

Query Processing over Data Streams

Continuous queries over continuous data streams

- Update time / stream element
- Query processing time

⇒ **Approximation & Randomization**

Methods

- Batch processing
- Sampling
 - Evaluate query over a sample of the stream
- Synopsis data structures
 - Fast update and fast query time

Class of Queries

Aggregates over Joins

- Equi-join J between relations F and G with the join constraint $F.a = G.a$
- Queries specified by a join and an aggregate: COUNT, SUM

Example

- Stream F :

a	1	1	2	3	1	3
---	---	---	---	---	---	---

, frequency vector $\mathbf{f}$:

i	1	2	3
f_i	3	1	2
- Stream G :

a	3	1	3	1	1
---	---	---	---	---	---

, frequency vector $\mathbf{g}$:

i	1	2	3
g_i	3	0	2

$$\begin{aligned}
 \text{COUNT}(F \bowtie_a G) &= \mathbf{f}\mathbf{g}^T \\
 &= [3 \quad 1 \quad 2] \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix} \\
 &= 3 \cdot 3 + 1 \cdot 0 + 2 \cdot 2 \\
 &= 13
 \end{aligned}$$

Basic Sketching Technique (AGMS99)

Main Idea:

- Summarize frequency table by projecting it onto a random vector $\Rightarrow$ *sketch*
- Use sketches to recover query result

Random vectors:

- $\xi = [\xi_1 \dots \xi_n]$ random vector of ± 1 values, called ξ **family**

Sketches:

- Sketch of F , $X_F = \mathbf{f}\xi^T$
- Sketch of G , $X_G = \mathbf{g}\xi^T$
- $X = X_F X_G$ estimates $\text{COUNT}(F \bowtie_a G)$ since

$$E[X] = E[\mathbf{f}\xi^T \xi \mathbf{g}^T] = \mathbf{f}E[\xi^T \xi]\mathbf{g}^T = \mathbf{f}I\mathbf{g}^T = \mathbf{f}\mathbf{g}^T$$

if $E[\xi^T \xi] = I$.

$\Rightarrow$ distinct elements of ξ must be pair-wise independent $\forall i \neq i', \xi_i^2 = 1, E[\xi_i \xi_{i'}] = 0$

Basic Sketching Technique

Example: $\xi = [\xi_1 \ \xi_2 \ \xi_3] = [-1 \ +1 \ -1]$

$$X_F = \mathbf{f}\xi^T = [3 \ 1 \ 2] \begin{bmatrix} -1 \\ +1 \\ -1 \end{bmatrix} = -4$$

$$X_G = \mathbf{g}\xi^T = [3 \ 0 \ 2] \begin{bmatrix} -1 \\ +1 \\ -1 \end{bmatrix} = -5$$

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

- Error of estimate X due to its variance

$$\text{Var}(X) \leq 2\mathbf{f}\mathbf{f}^T \mathbf{g}\mathbf{g}^T = 2 \text{SJ}(F) \text{SJ}(G)$$

- Only properties of ξ needed: $\forall i_1 \neq i_2 \neq i_3 \neq i_4, E[\xi_{i_1} \xi_{i_2}] = 0, E[\xi_{i_1} \xi_{i_2} \xi_{i_3} \xi_{i_4}] = 0$
- ξ need not have elements fully independent. 4-wise independence suffices
 - $\Rightarrow \xi$ can be generated from **small seeds** using $\xi_i(s) = h(s, i)$ [Alon et al. 86]
 - $\Rightarrow \xi$ vector is not stored. Required elements generated on the fly from seeds

Sketch Maintenance over Streams

- Choose seed s that generates ξ

- Stream F :

a	1	1	2	3	1	3
---	---	---	---	---	---	---

, frequency vector $\mathbf{f}$:

i	1	2	3
f_i	3	1	2

$$\begin{aligned} X_F &= \mathbf{f}\xi^T = \sum_i f_i \xi_i = \sum_i (\xi_i + \dots + \xi_i) \\ &= \sum_{t \in F} \xi_{t.a} = \xi_1 + \xi_1 + \xi_2 + \xi_3 + \xi_1 + \xi_3 \\ &= h(s, 1) + h(s, 1) + h(s, 2) + h(s, 3) + h(s, 1) + h(s, 3) \end{aligned}$$

- Stream G :

a	3	1	3	1	1
---	---	---	---	---	---

, frequency vector $\mathbf{g}$:

i	1	2	3
g_i	3	0	2

$$\begin{aligned} X_G &= \mathbf{g}\xi^T = \sum_i g_i \xi_i = \sum_i (\xi_i + \dots + \xi_i) \\ &= \sum_{t \in G} \xi_{t.a} = \xi_3 + \xi_1 + \xi_3 + \xi_1 + \xi_1 \\ &= h(s, 3) + h(s, 1) + h(s, 3) + h(s, 1) + h(s, 1) \end{aligned}$$

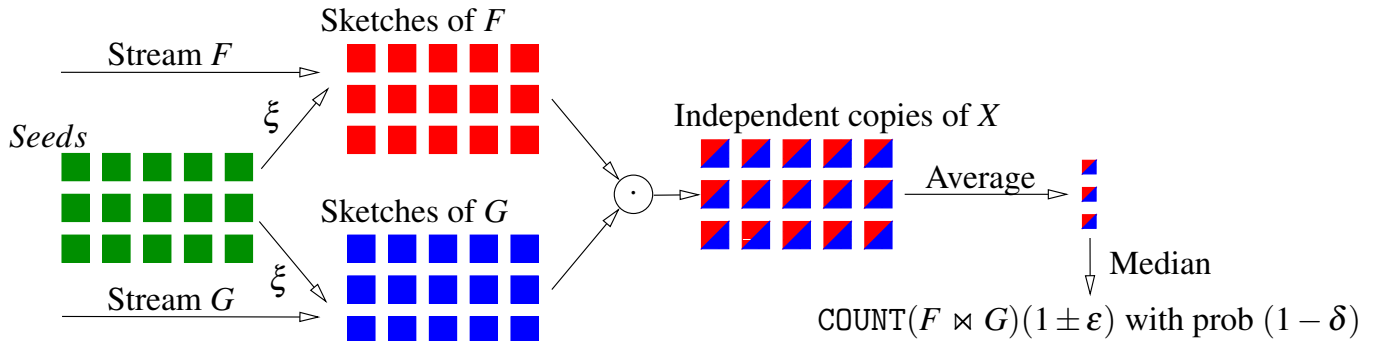
- Counters X_F and X_G need only log space in the size of the stream

Sketch Error Reduction

- Estimation of $\text{COUNT}(F \bowtie_a G)$ from single sketches of F and G is *too noisy*

Solution:

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



- Memory required independent of the size of the stream

Contributions

- Pseudo-random number generating schemes (SIGMOD '06, TODS '07)
 - EH3 (3-wise independent) is fast and most accurate
- Sketches over samples (under review)
 - Similar accuracy for 1% sample
 - A factor of 100 speed improvement
- Statistical analysis of sketch estimators (SIGMOD '07, under review)
 - Fast-AGMS sketches are the most accurate and fast enough
- Sketches for interval data (SIGMOD '06, TODS '07, under review)
 - Fast range-summation is more accurate
 - DMAP is faster

Outline of the Talk

- Pseudo-random number generating schemes
 - Limited Independence Random Variables
- Sketches over samples
- Statistical analysis of sketch estimators
- Sketches for interval data

4-Wise Independent Random Variables

Problem

- Domain $I = \{0, 1, \dots, 2^n - 1\}$, $n \geq 0$
- 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

Scheme Comparison

Scheme	Time (ns)	Seed size
BCH3	10.8	$n + 1$
EH3	7.3	$n + 1$
Polynomials2	31.4	$2n$
Tabulation2	5.1	$2^b \cdot \frac{n}{b}$
BCH5	12.7	$2n + 1$
Polynomials4	106.4	$4n$
RM7	3,301	$1 + n + \frac{n(n-1)}{2}$
Tabulation4	10.3	$2^b \cdot \left(\frac{n}{b}\right)^{\log_2 3}$

- n is the number of bits
- b represents the size (number of bits) of the stored sub-sequences

Basic Sketching Technique Variance

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

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

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

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

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

BCH3 Scheme (ABI 1986)

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

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

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

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

EH3 Scheme (FKSV 2002)

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

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

$$= \begin{cases} 0 & \text{if } i \oplus j \oplus k \oplus l \neq \bar{0} \\ 1 & \text{if } (i \oplus j \oplus k \oplus l = \bar{0}) \wedge (h(i) \oplus h(j) \oplus h(k) \oplus h(l) = 0) \\ -1 & \text{if } (i \oplus j \oplus k \oplus l = \bar{0}) \wedge (h(i) \oplus h(j) \oplus h(k) \oplus h(l) = 1) \end{cases}$$

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

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

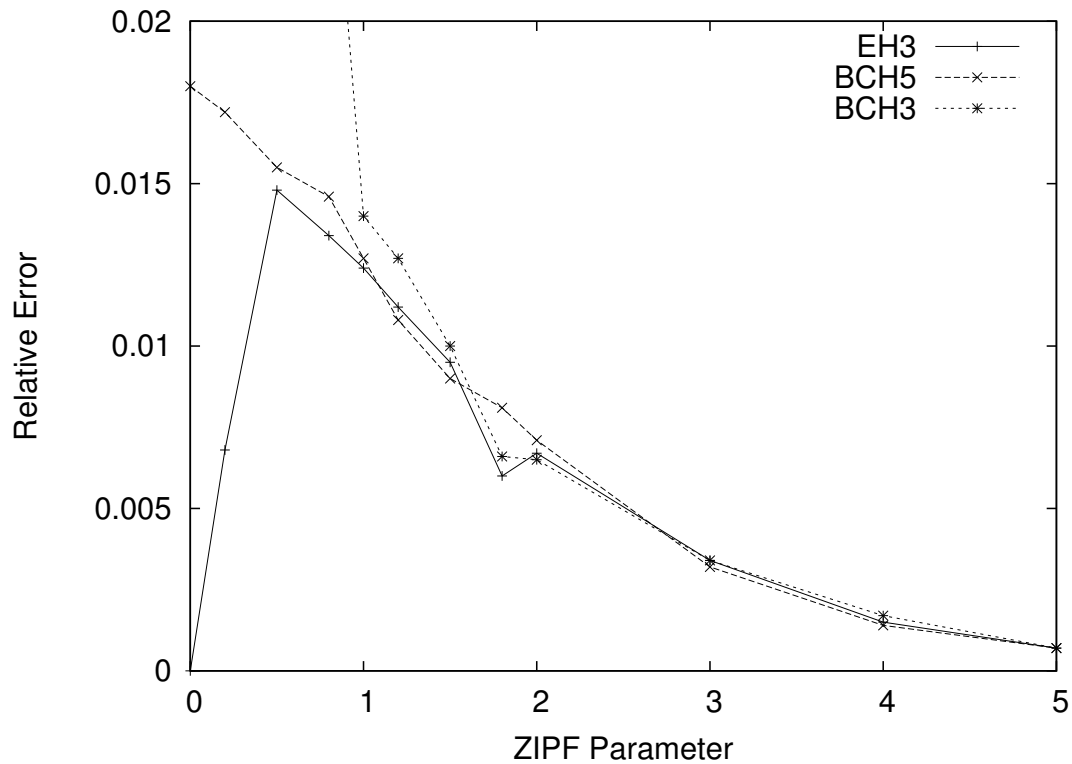
- Best case analysis for $f_i = f$, $g_j = g$, $|I| = 4^n$

$$\textcolor{green}{\text{Var} [\text{EH3}] = 0}$$

BCH3 vs EH3 vs BCH5

Settings

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



Outline of the Talk

- Pseudo-random number generating schemes
- Sketches over samples
 - Combination of two basic estimators
- Statistical analysis of sketch estimators
- Sketches for interval data

Sketches over Sampled Streams

• 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	1	1	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

• $\xi = [\xi_1 \ \xi_2 \ \xi_3] = [-1 \ +1 \ -1]$

$$X'_F = \mathbf{f}' \xi^T = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ +1 \\ -1 \end{bmatrix} = -1$$

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

$$X = C \cdot X'_F X'_G = C \cdot (-1)(-2) = C \cdot 2$$

Generic Sampling

- Sample at the tuple (stream element) level
- Frequency vector is a vector of random variables

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

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

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

Sketches over Samples

- Build the sketch over the sample (not materialized)

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

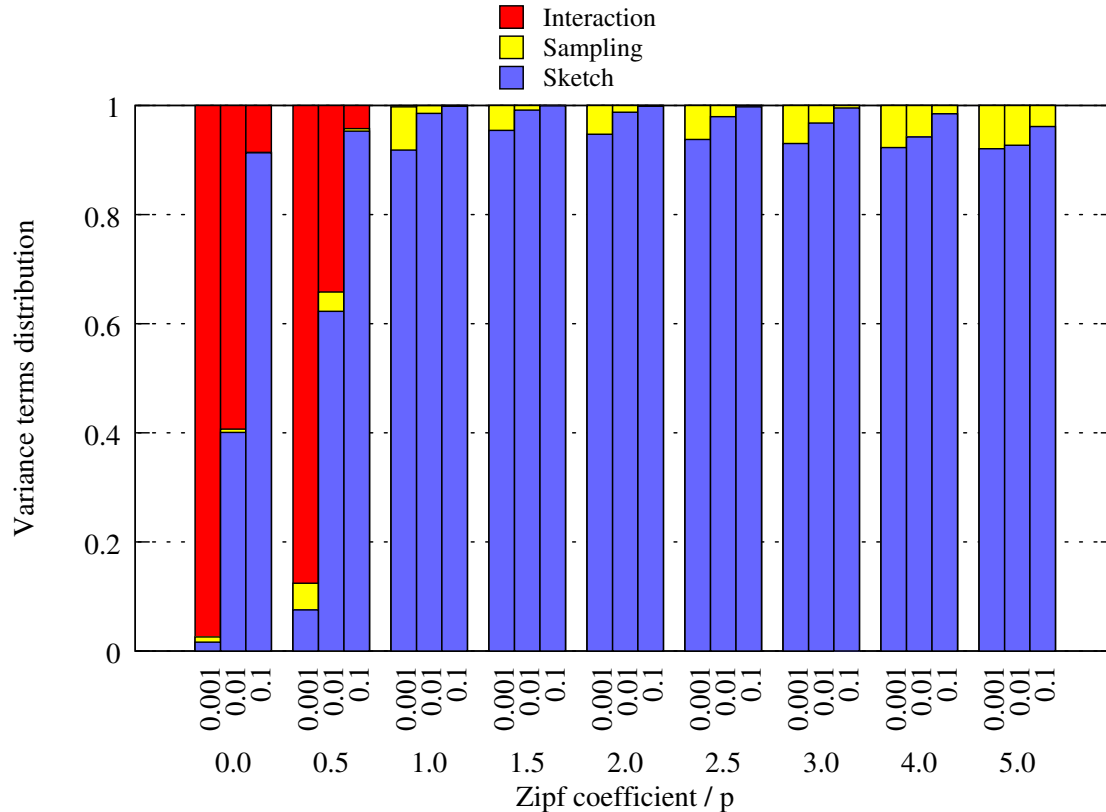
$$E[X] = C \cdot \sum_{i \in I} E[f'_i] E[g'_i]$$

$$\begin{aligned} \text{Var}[X] = & C^2 \cdot \left[\sum_{i \in I} E[f_i'^2] \sum_{j \in I} E[g_j'^2] + 2 \cdot \sum_{i \in I} \sum_{j \in I} E[f_i' f_j'] E[g_i' g_j'] \right. \\ & \left. - 2 \cdot \sum_{i \in I} E[f_i'^2] E[g_i'^2] - \left(\sum_{i \in I} E[f_i'] E[g_i'] \right)^2 \right] \end{aligned}$$

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

Variance for Bernoulli Sampling

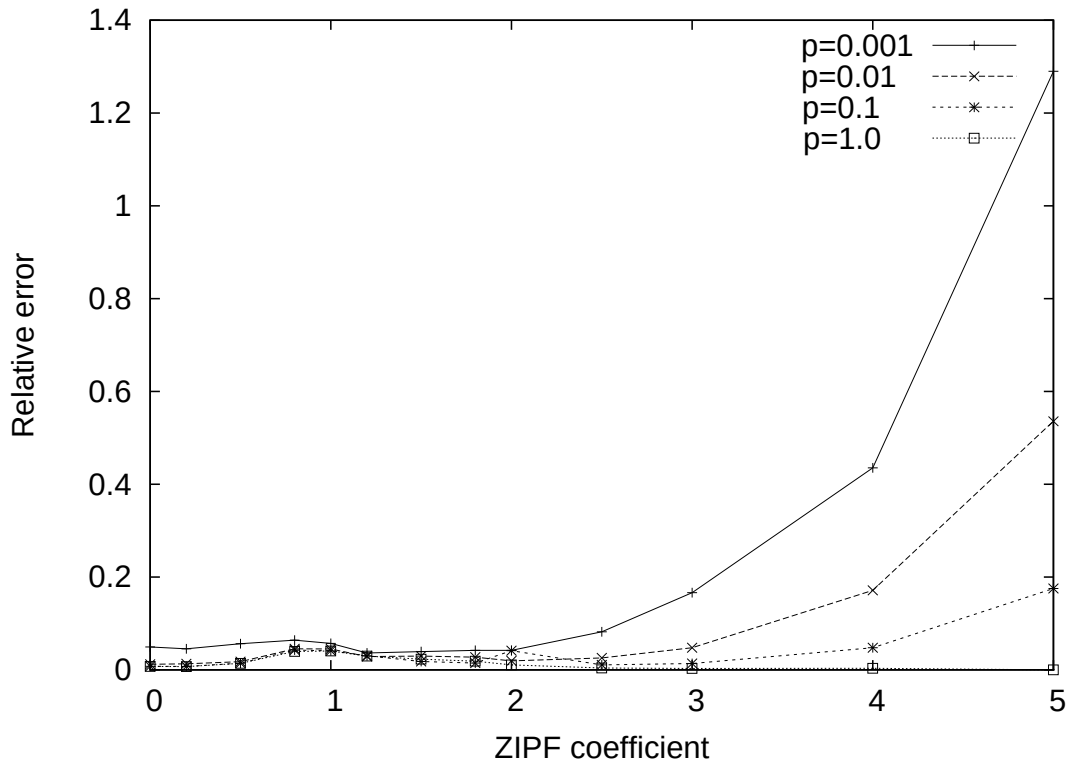
- Term significance as a function of the frequency distribution



Error for Bernoulli Sampling

Settings

- Domain size = 1 million, Tuples = 100 millions, independent relations



Outline of the Talk

- Pseudo-random number generating schemes
- Sketches over samples
- Statistical analysis of sketch estimators
 - Improve the confidence bounds (distribution-dependent)
- Sketches for interval data

Sketches

Idea

- Summarize frequency vector in small space $\Rightarrow$ approximate results

Definition

- Randomized data structure $\Rightarrow$ matrix of counters $x = (x_{ij})$
 - Randomization, update
- Size of join estimator
 - Error, confidence intervals (bounds)

Types

- AGMS
- Fast-AGMS (F-AGMS)
- Fast-Count (FC)
- Count-Min (CM)

Sketch Comparison

Big- $\mathcal{O}$

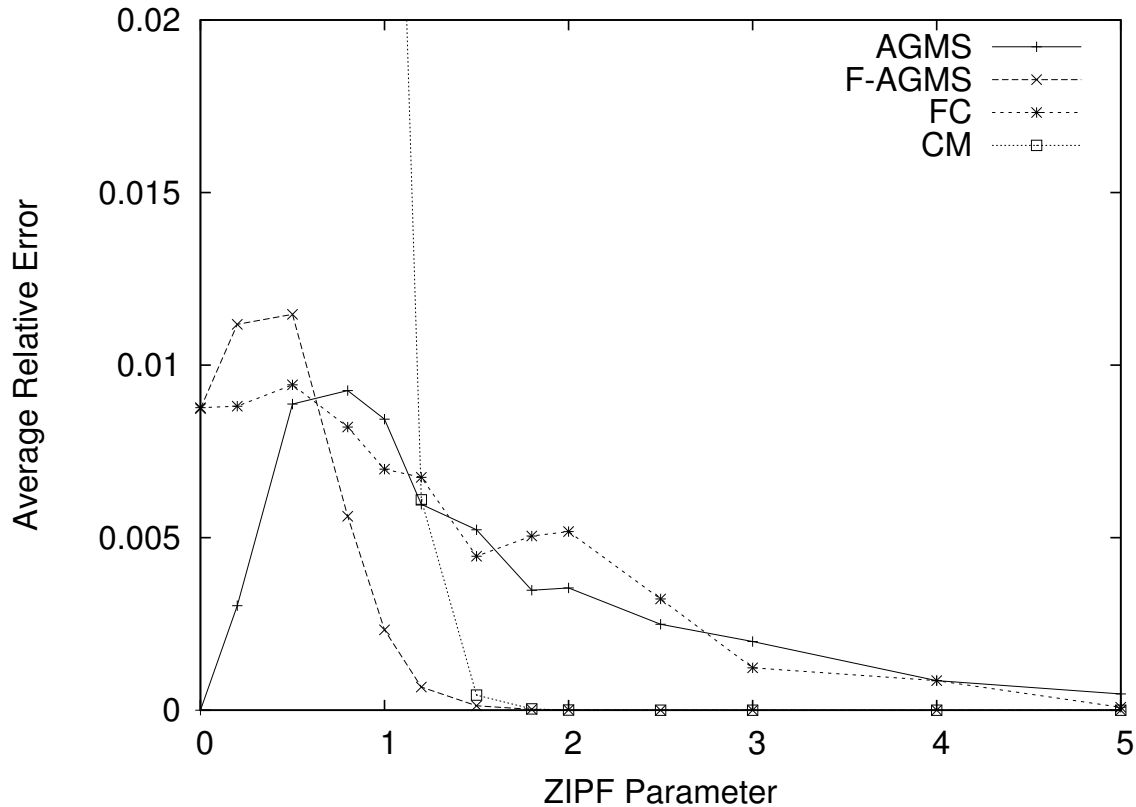
Sketch	Size	Update Time	Error
AGMS	$\mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$	$\mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$	$Z \in (\bar{f} \odot \bar{g} \pm \varepsilon \ \bar{f}\ _2 \ \bar{g}\ _2)$
F-AGMS	$\mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$	$\mathcal{O}\left(\log \frac{1}{\delta}\right)$	$Z \in (\bar{f} \odot \bar{g} \pm \varepsilon \ \bar{f}\ _2 \ \bar{g}\ _2)$
FC	$\mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$	$\mathcal{O}\left(\log \frac{1}{\delta}\right)$	$Z \in (\bar{f} \odot \bar{g} \pm \varepsilon \ \bar{f}\ _2 \ \bar{g}\ _2)$
CM	$\mathcal{O}\left(\frac{1}{\varepsilon} \cdot \log \frac{1}{\delta}\right)$	$\mathcal{O}\left(\log \frac{1}{\delta}\right)$	$\bar{f} \odot \bar{g} \leq Z \leq \bar{f} \odot \bar{g} \pm \varepsilon \ \bar{f}\ _1 \ \bar{g}\ _1$

Goals

- Reduce randomization
- Maintain (improve) error bounds

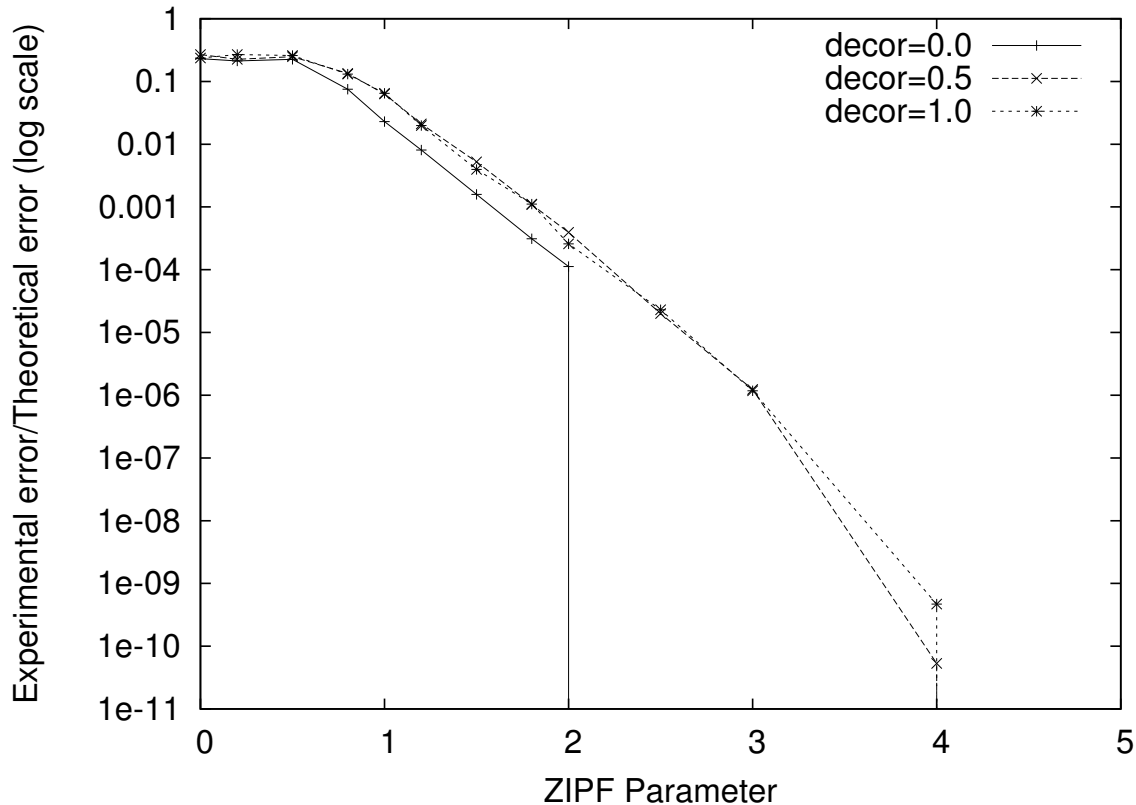
Experimental Comparison

Setup: self-join size, same sketch size, relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)



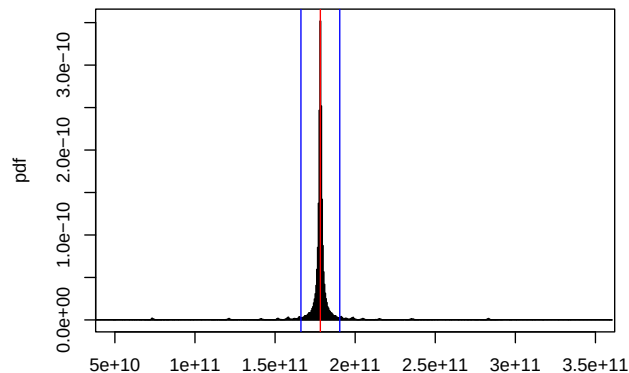
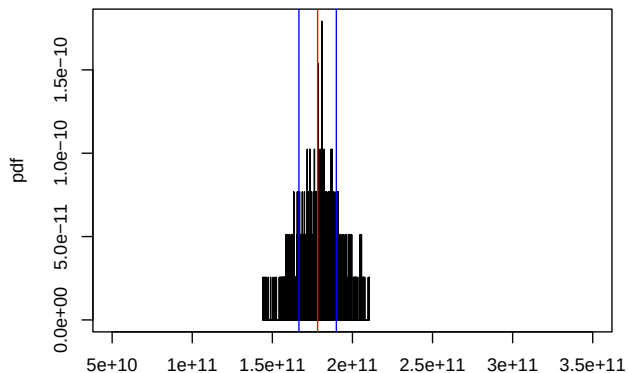
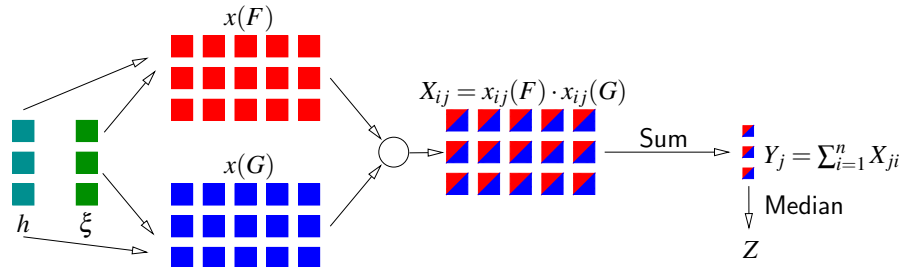
Confidence Bounds (F-AGMS)

Setup: size of join with different correlations, 95% confidence bounds

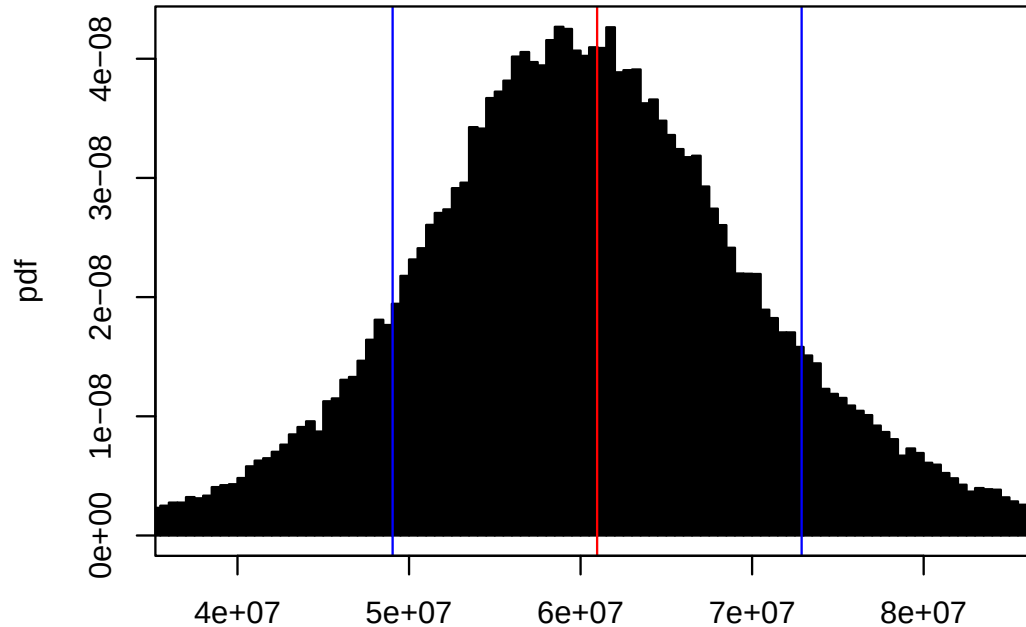


Fast-AGMS Sketches (CG05)

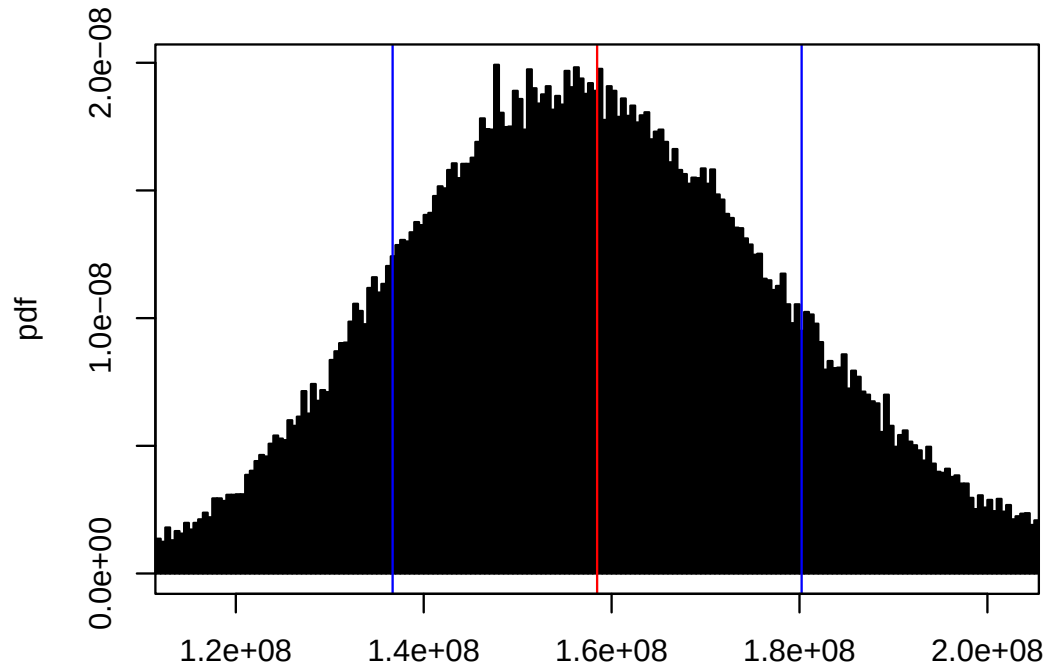
Setup: self-join size, Zipf=1.5, true result $\approx 1.8 \cdot 10^{11}$



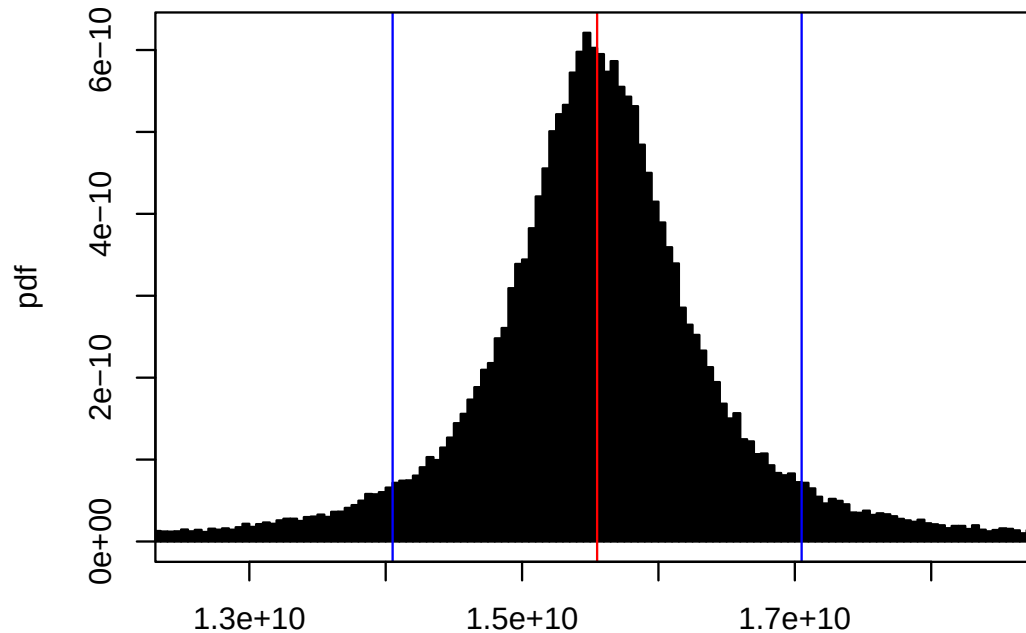
Zipf=0.0



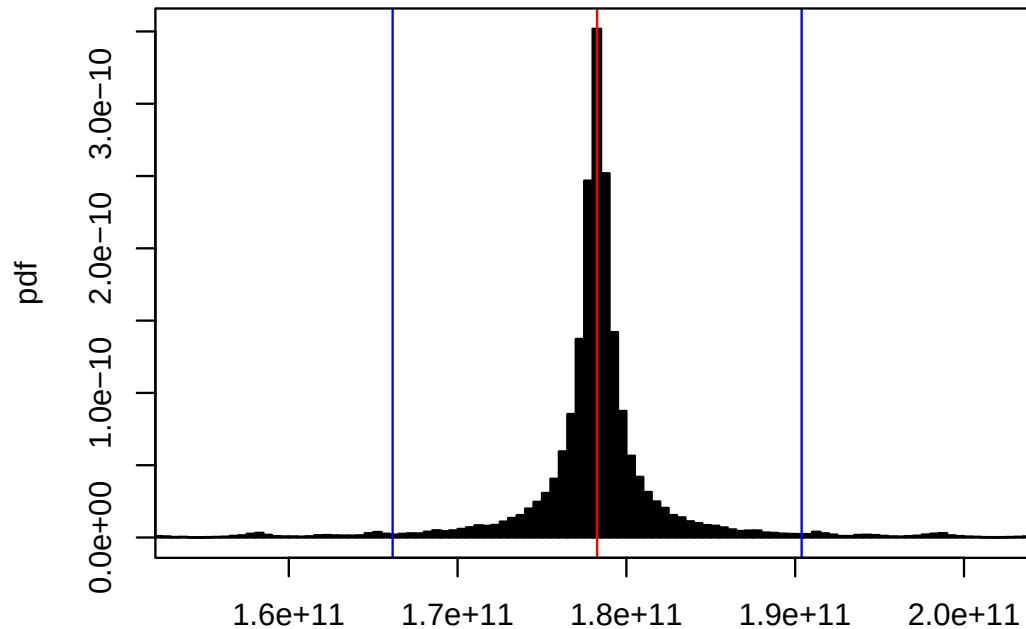
Zipf=0.5



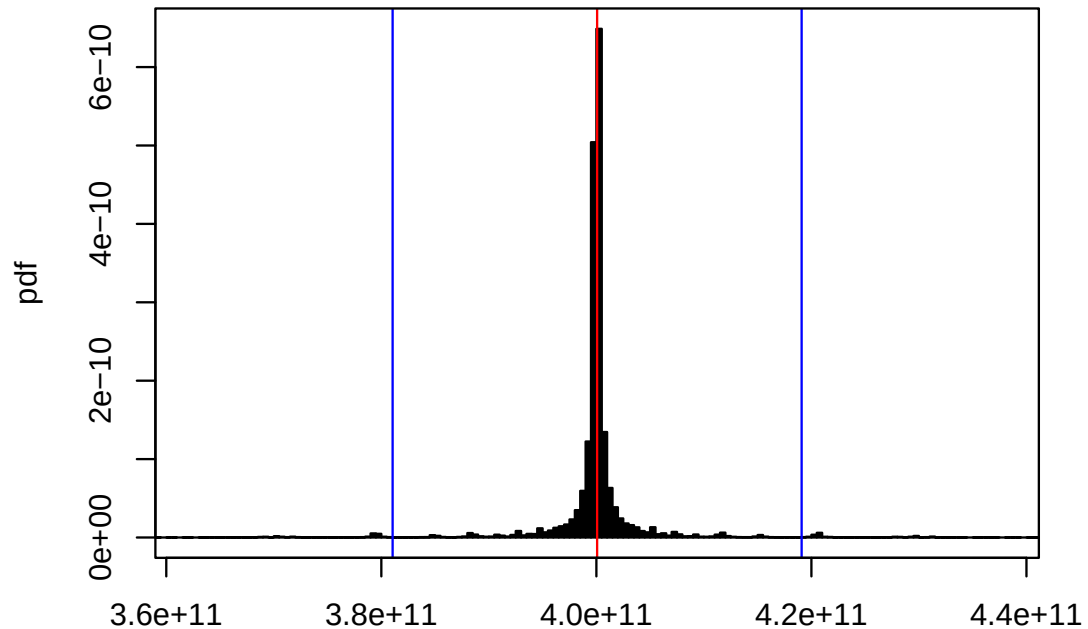
Zipf=1.0



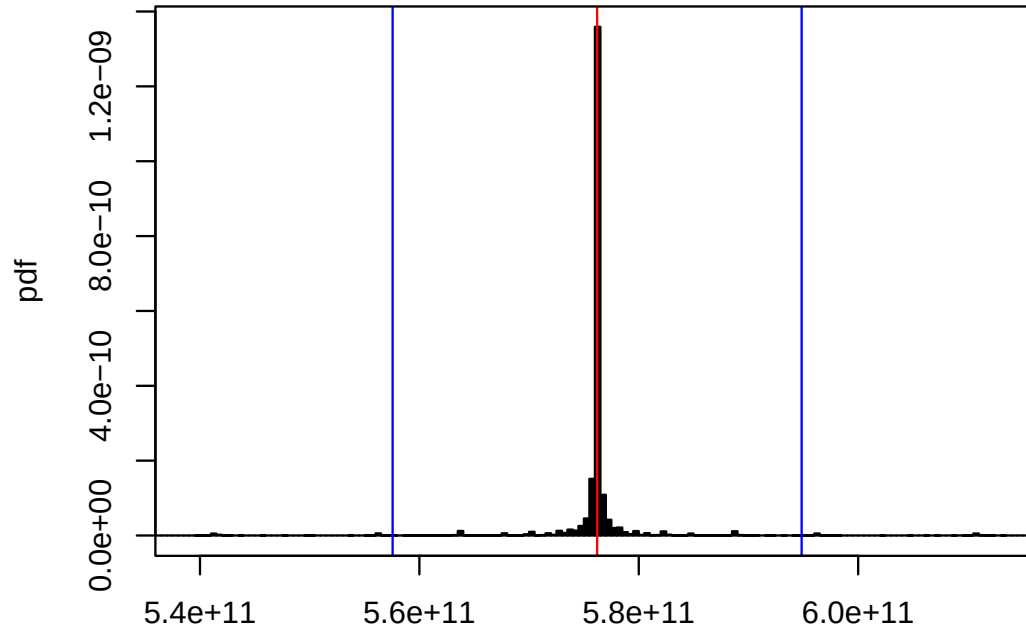
Zipf=1.5



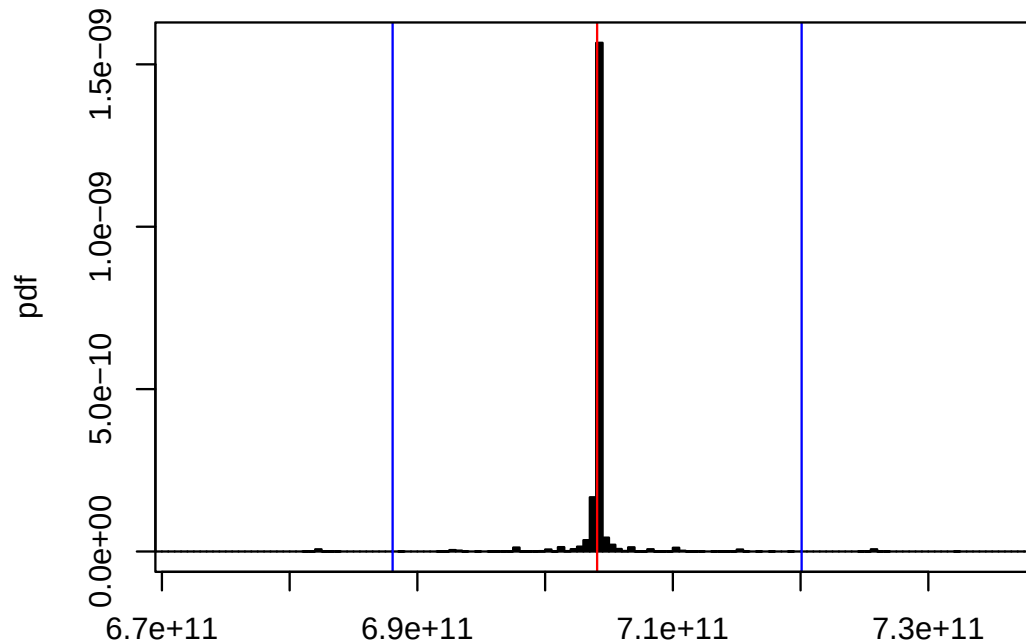
Zipf=2.0



Zipf=2.5



Zipf=3.0



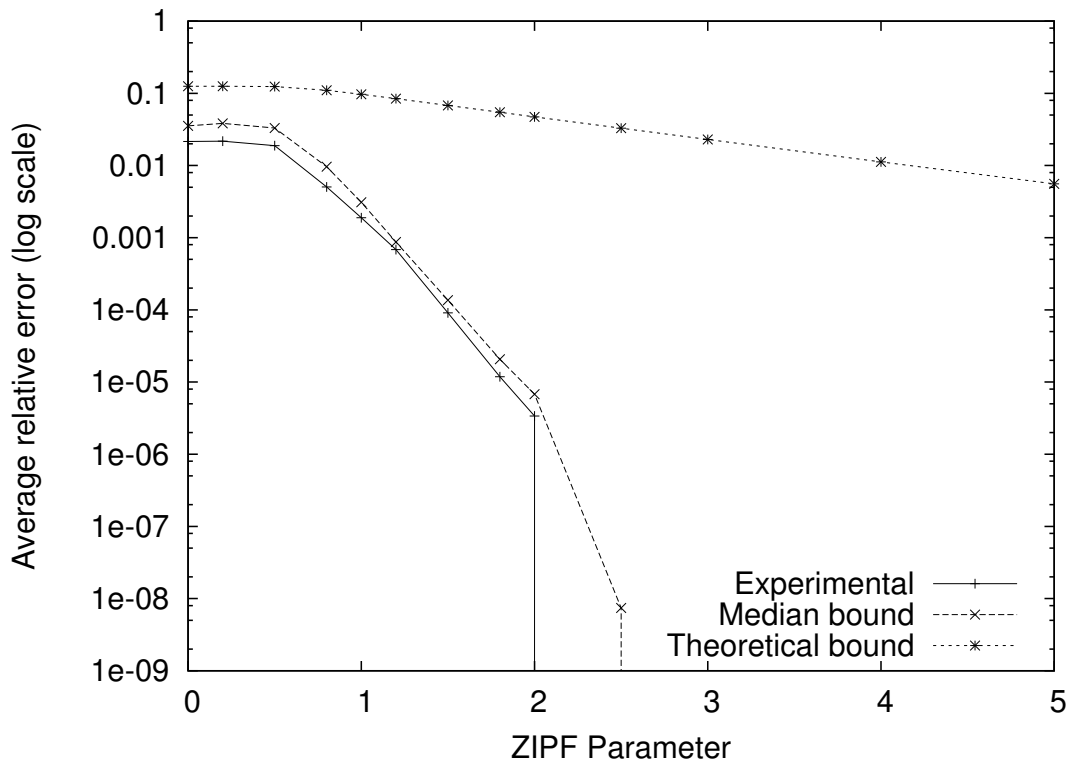
Fast-AGMS Sketches (CG05)

Confidence bounds

- Frequency moments of X : $E[X]$, $\text{Var}[X]$
- Estimator for true value: median Z
- Distribution-independent
 - Chebyshev & Chernoff bounds
 - $Z \in (\bar{f} \odot \bar{g} \pm \epsilon \|\bar{f}\|_2 \|\bar{g}\|_2)$ with probability at least $1 - \delta$
 - Identical to AGMS sketches (expectation and variance are the same)
- Distribution-dependent
 - Median CLT
 - Asymptotic distribution of Z is Student t

Fast-AGMS Sketches (CG05)

Setup: self-join size, relative error ($\frac{|\text{actual} - \text{estimate}|}{\text{actual}}$), 95% confidence bounds



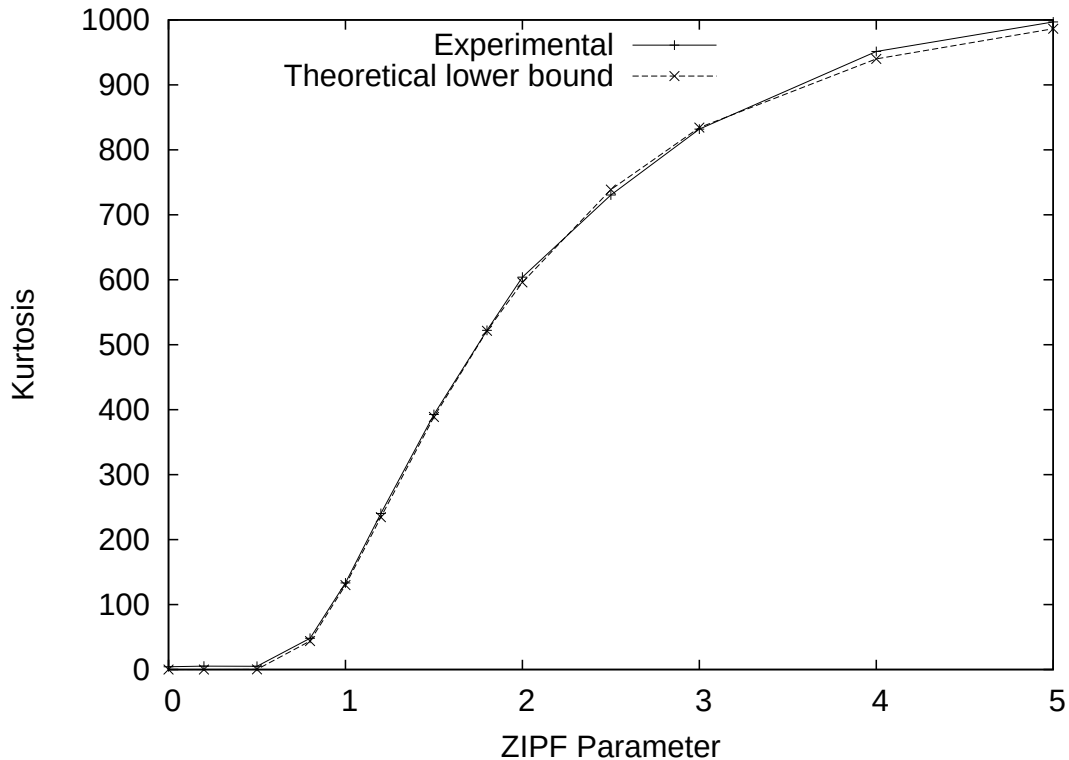
AGMS vs Fast-AGMS

Mean vs Median

- Efficiency: $e(f) = 4f(\theta)^2 \text{Var}[X]$
 - $e(f) > 1 \Rightarrow \text{Median}$, $e(f) < 1 \Rightarrow \text{Mean}$
- Kurtosis: $k = \frac{E[(X - E[X])^4]}{\text{Var}[X]^2}$
 - $k > 6 \Rightarrow \text{Median}$, $k \leq 6 \Rightarrow \text{Mean}$
 - $k = 3$ for normal distribution
- No quantitative relationship between efficiency and kurtosis
- Qualitatively high kurtosis implies high efficiency

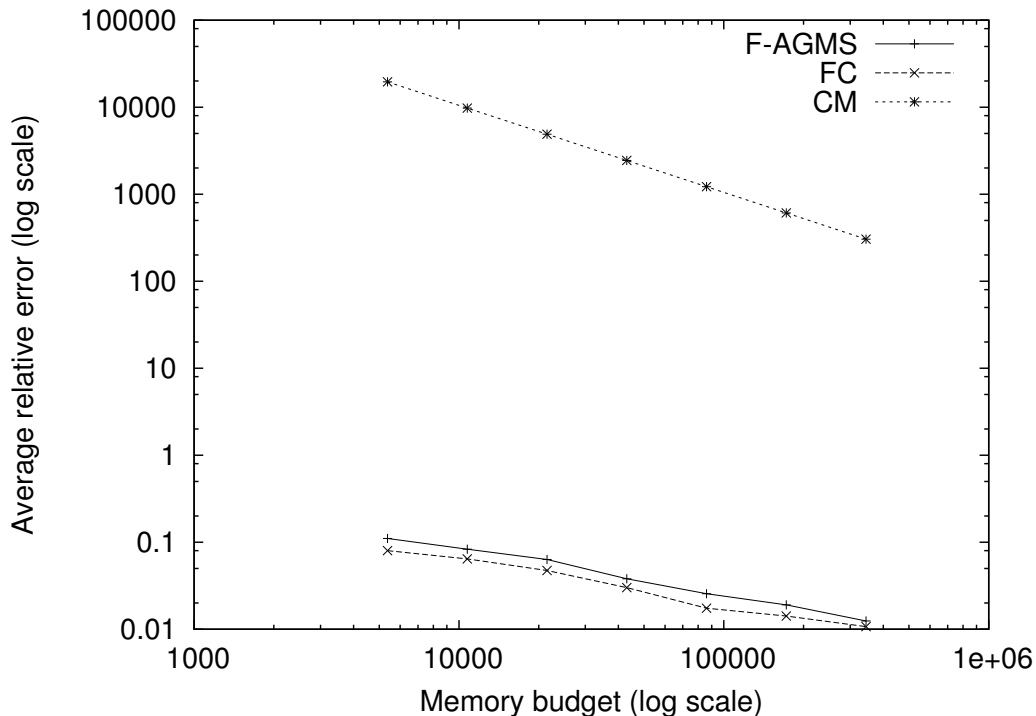
Fast-AGMS Kurtosis

Setup: self-join size, lower approximation of kurtosis $Var_h[Var_\xi[X]]$



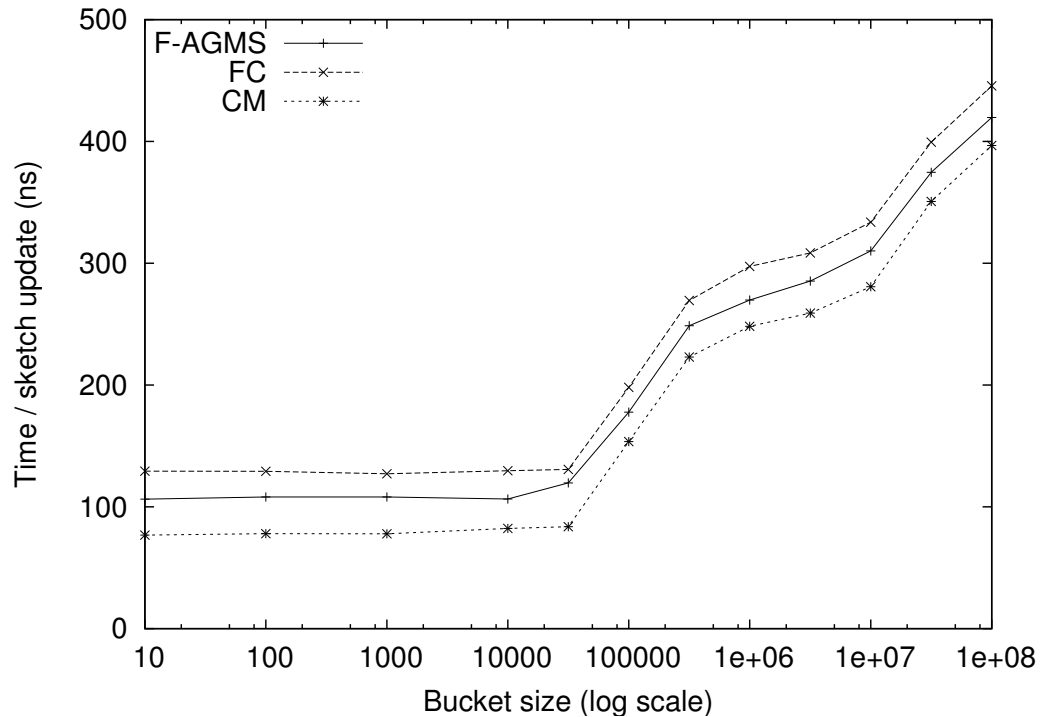
Memory Budget ((EN06) memory unpeaked)

Setup: size of join, sketch size (row=21), relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)



Update Time

Setup: sketch size (row=1), Xeon 2.8 GHz, 512 KB cache, 4 GB main memory



Outline of the Talk

- Pseudo-random number generating schemes
- Sketches over samples
- Statistical analysis of sketch estimators
- Sketches for interval data
 - Fast range-sumable random variables

Class of Applications

Spatial Size of Join, Selectivity Estimation

- Reduction to [Size of Join](#)
- Solution based on [AGMS 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
- Dyadic Mapping DMAP [GKMS 2002, CM 2004, DGR 2004]
- Fast Range-Summation [GKMS 2003, CGLMS 2005]

Derived Problem

Sketches for interval data

- 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 $|F \bowtie_a G|$

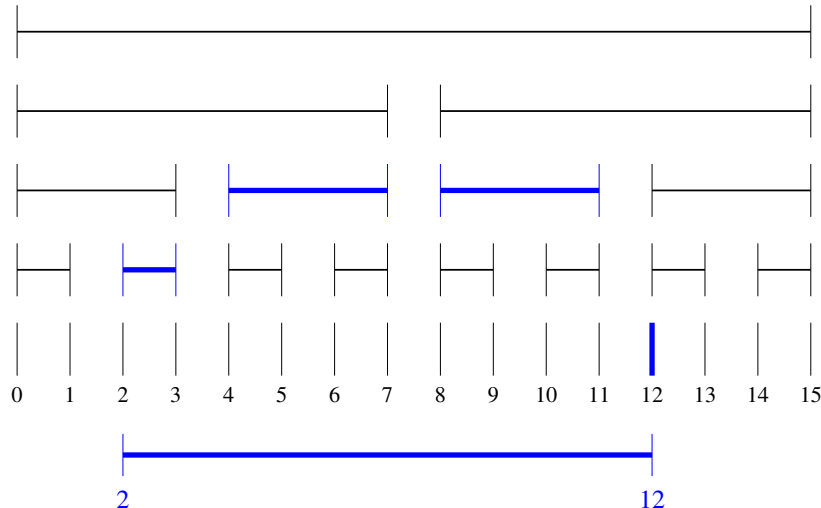
Dyadic Mapping DMAP (1)

Dyadic Intervals [GKMS 2002]

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

Minimal Dyadic Cover

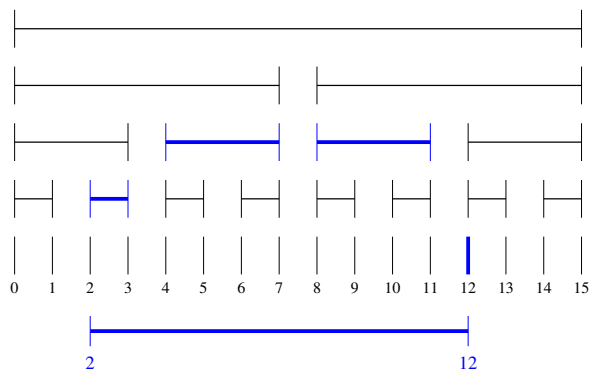
- $D([\alpha, \beta)) = [\alpha = \alpha_1, \beta_1) \cup [\alpha_2, \beta_2) \cup \dots \cup [\alpha_k, \beta_k = \beta)$
- $|D([\alpha, \beta))|$ is $O(\log(\beta - \alpha))$



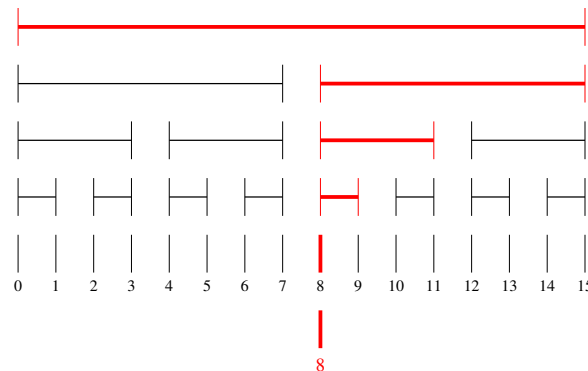
Dyadic Mapping DMAP (2)

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

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

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

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

4-Wise Independent Fast Range-Summable Random Variables

Problem

- $I = \{0, 1, \dots, 2^n - 1\}$, $n \geq 0$
- $\xi\{I\}$, $\xi_i(S) = (-1)^{f(s,i)}$
- 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-sumnable 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

EH3 is Fast Range-Summable

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

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

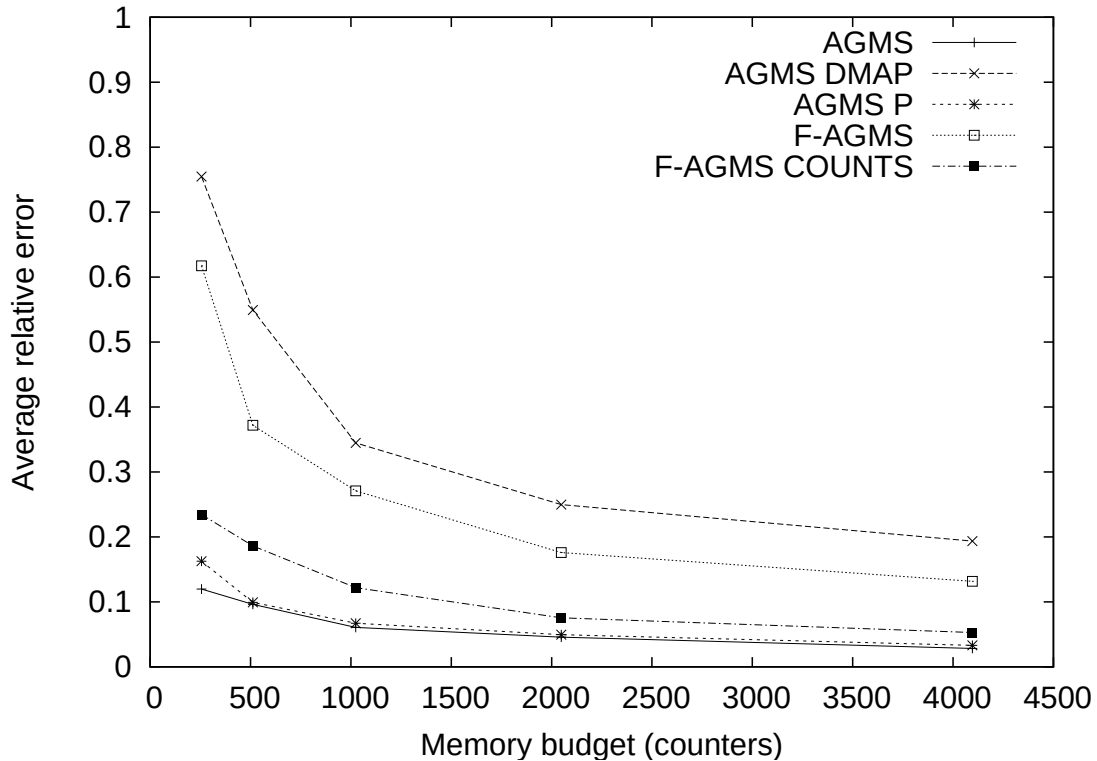
Sketching Intervals

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

Size of Spatial Join

Settings

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



Conclusions

- Pseudo-random number generating schemes (SIGMOD '06, TODS '07)
 - Use EH3 generating scheme
- Sketches over samples (under review)
 - Compute sketch over a sample
- Statistical analysis of sketch estimators (SIGMOD '07, under review)
 - Use Fast-AGMS sketches
- Sketches for interval data (SIGMOD '06, TODS '07, under review)
 - Use fast range-summation
- Sketch framework available online

⇒ Fast-AGMS sketches with EH3 random variables over a sample

Questions