

Statistical Analysis of Sketch Estimators

Florin Rusu and Alin Dobra
CISE Department
University of Florida

June 12, 2007

Size of Join over Data-Streams

Input Data

- Stream F : (1,5) (1,2) (2,3) (3,1) (1,-3) (3,2)

$$\bar{\mathbf{f}}: \begin{array}{c|ccc} \text{key} & 1 & 2 & 3 \\ \hline \text{freq} & 4 & 3 & 3 \end{array}$$

- Stream G : (2,1) (3,2) (1,3) (3,-2) (1,2) (1,2)

$$\bar{\mathbf{g}}: \begin{array}{c|ccc} \text{key} & 1 & 2 & 3 \\ \hline \text{freq} & 7 & 1 & 0 \end{array}$$

- Massive, high speed

Problem

- Size of Join $|F \bowtie G| = \bar{\mathbf{f}} \odot \bar{\mathbf{g}}$

$$|F \bowtie G| = \bar{\mathbf{f}} \odot \bar{\mathbf{g}} = [4 \ 3 \ 3] \cdot \begin{bmatrix} 7 \\ 1 \\ 0 \end{bmatrix} = 4 \cdot 7 + 3 \cdot 1 + 3 \cdot 0 = 31$$

- Small space, single pass

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)

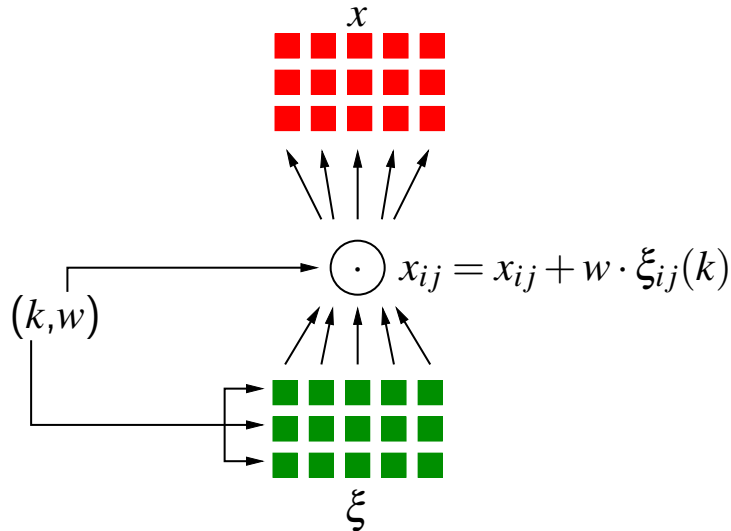
AGMS Sketches (AMS96,AGMS99)

Randomization

- Matrix $\xi = (\xi_{ij})$ of 4-wise independent ± 1 random variables, $\xi_{ij} : I \rightarrow \{-1, +1\}$

Update

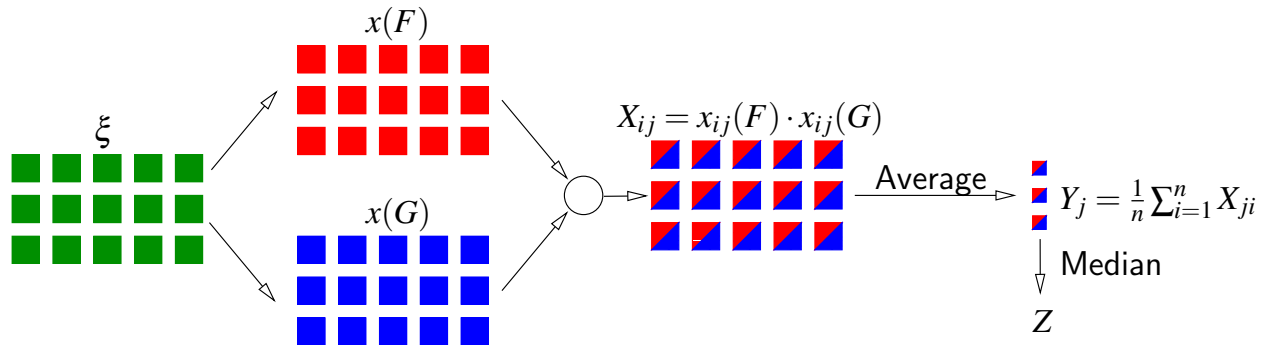
- Time \approx sketch size



AGMS Sketches (AMS96,AGMS99)

Size of Join Estimator

- $E[Z] = \bar{f} \odot \bar{g} = |F \bowtie G|$, $Z \in (\bar{f} \odot \bar{g} \pm \epsilon \|\bar{f}\|_2 \|\bar{g}\|_2)$ with probability at least $1 - \delta$
- $\|\bar{f}\|_2 = \sqrt{\bar{f} \odot \bar{f}} = \sqrt{\sum_{i \in I} f_i^2}$, $\|\bar{g}\|_2 = \sqrt{\bar{g} \odot \bar{g}} = \sqrt{\sum_{i \in I} g_i^2}$
- Sketch size: $n = \mathcal{O}(\frac{1}{\epsilon^2})$ and $m = \mathcal{O}(\log \frac{1}{\delta})$



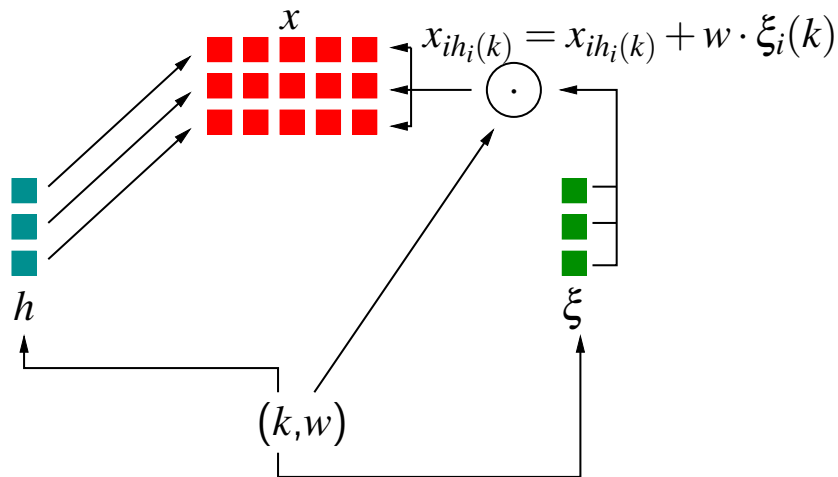
Fast-AGMS Sketches (CG05)

Randomization

- Vector h_i of 2-universal hash functions, $h_i : I \rightarrow B$
- Vector ξ_i of 4-wise independent ± 1 random variables, $\xi_i : I \rightarrow \{-1, +1\}$

Update

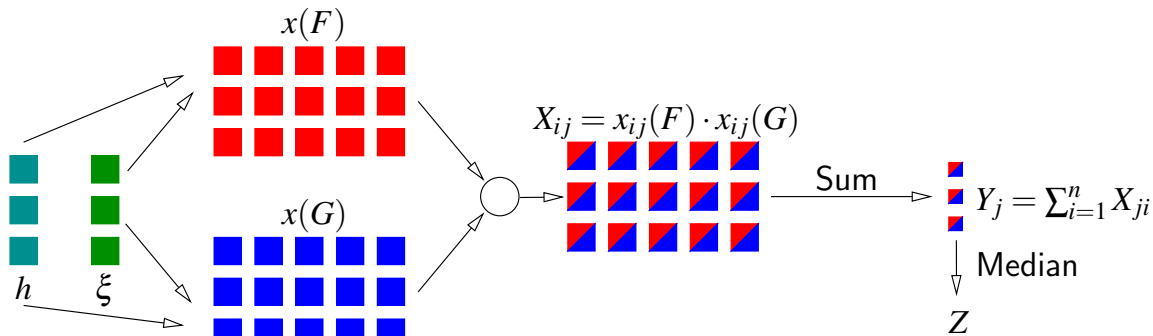
- Time \approx no. of rows m



Fast-AGMS Sketches (CG05)

Size of Join Estimator

- $E[Z] = \bar{f} \odot \bar{g} = |F \bowtie G|$, $Z \in (\bar{f} \odot \bar{g} \pm \epsilon \|\bar{f}\|_2 \|\bar{g}\|_2)$ with probability at least $1 - \delta$
- $\|\bar{f}\|_2 = \sqrt{\bar{f} \odot \bar{f}} = \sqrt{\sum_{i \in I} f_i^2}$, $\|\bar{g}\|_2 = \sqrt{\bar{g} \odot \bar{g}} = \sqrt{\sum_{i \in I} g_i^2}$
- Sketch size: $B = n = \mathcal{O}(\frac{1}{\epsilon^2})$ and $m = \mathcal{O}(\log \frac{1}{\delta})$



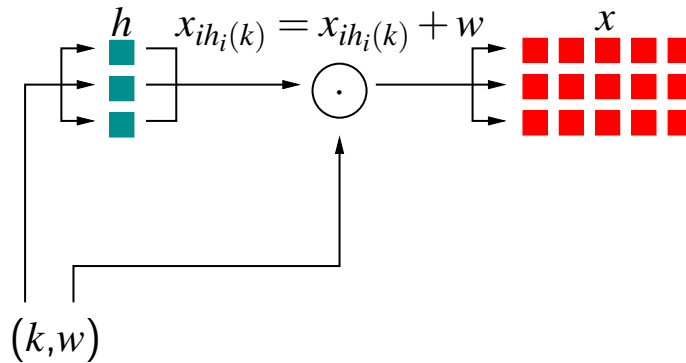
Fast-Count Sketches (TZ04)

Randomization

- Vector h_i of 4-universal hash functions, $h_i : I \rightarrow B$

Update

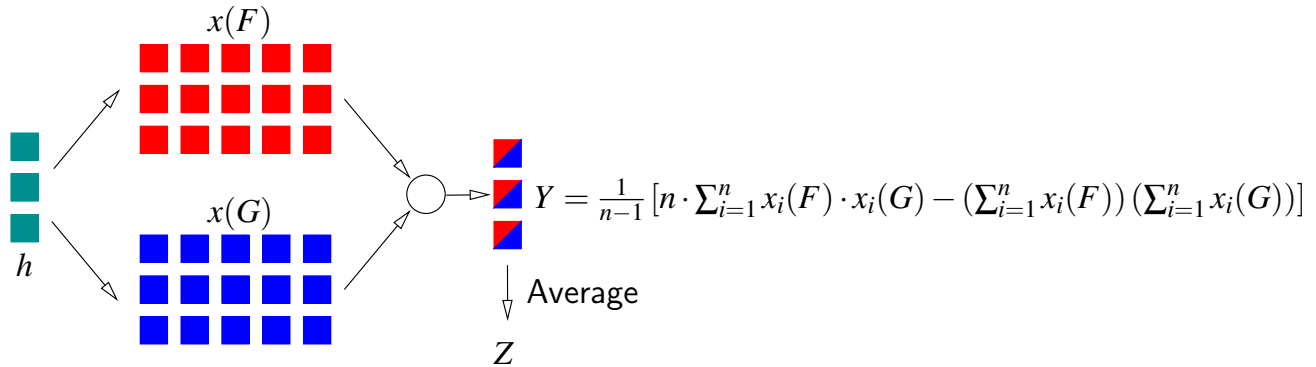
- Time \approx no. of rows m



Fast-Count Sketches (TZ04)

Size of Join Estimator

- $E[Z] = \bar{f} \odot \bar{g} = |F \bowtie G|$, $Z \in (\bar{f} \odot \bar{g} \pm \epsilon \|\bar{f}\|_2 \|\bar{g}\|_2)$ with probability at least $1 - \delta$
- $\|\bar{f}\|_2 = \sqrt{\bar{f} \odot \bar{f}} = \sqrt{\sum_{i \in I} f_i^2}$, $\|\bar{g}\|_2 = \sqrt{\bar{g} \odot \bar{g}} = \sqrt{\sum_{i \in I} g_i^2}$
- Sketch size: $B = n = \mathcal{O}(\frac{1}{\epsilon^2})$ and $m = \mathcal{O}(\log \frac{1}{\delta})$



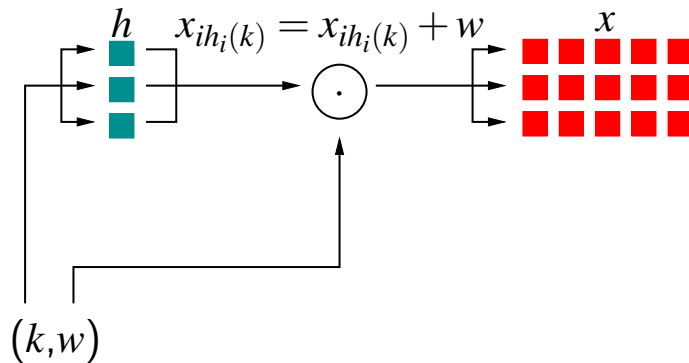
Count-Min Sketches (CM05)

Randomization

- Vector h_i of 2-universal hash functions, $h_i : I \rightarrow B$

Update

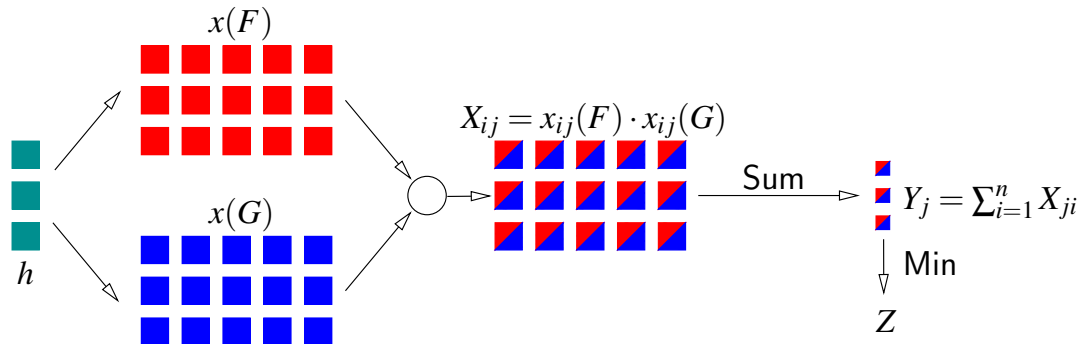
- Time \approx no. of rows m



Count-Min Sketches (CM05)

Size of Join Estimator

- $\bar{f} \odot \bar{g} \leq Z \leq \bar{f} \odot \bar{g} + \varepsilon \|\bar{f}\|_1 \|\bar{g}\|_1$ with probability at least $1 - \delta$ (over-estimate)
- $\|\bar{f}\|_1 = \sum_{i \in I} f_i$, $\|\bar{g}\|_1 = \sum_{i \in I} g_i$
- Sketch size: $B = n = \mathcal{O}(\frac{1}{\varepsilon})$ and $m = \mathcal{O}(\log \frac{1}{\delta})$



Sketch Comparison

Big- \mathcal{O}

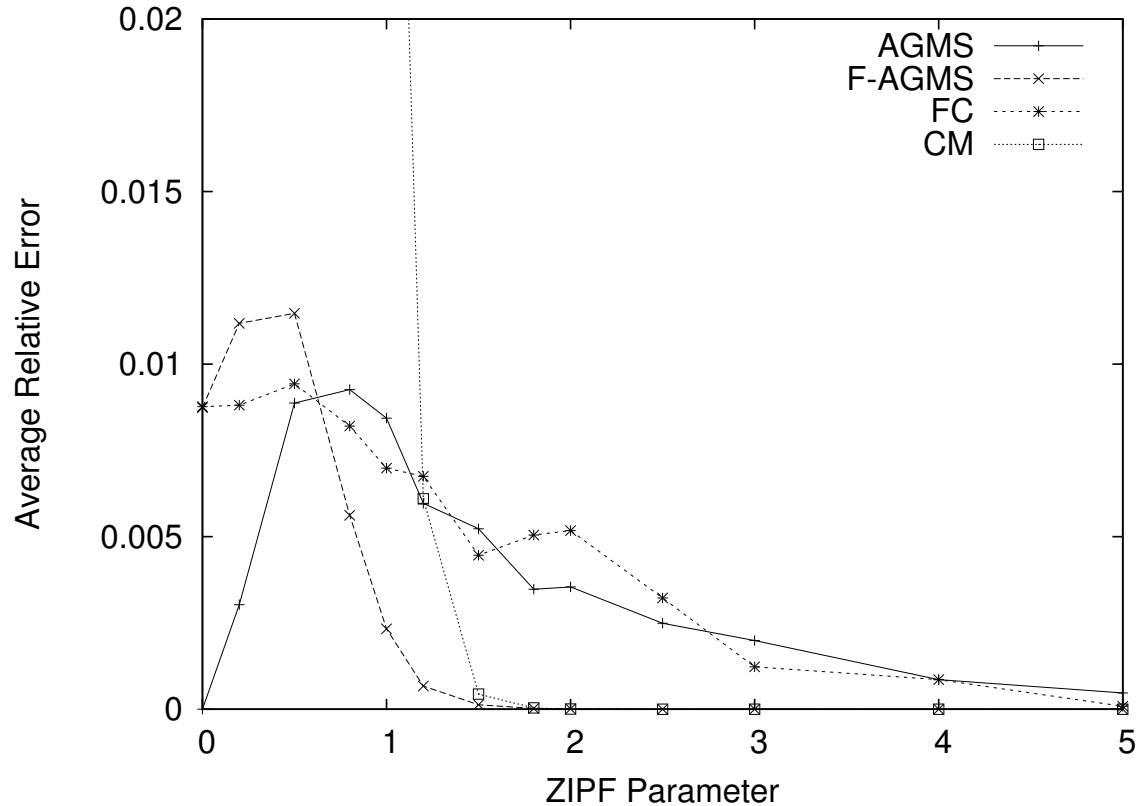
Sketch	Size	Update Time	Error
AGMS	$\mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$	$\mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$	$Z \in (\bar{f} \odot \bar{g} \pm \varepsilon \ \bar{f}\ _2 \ \bar{g}\ _2)$
F-AGMS	$\mathcal{O}\left(\frac{1}{\varepsilon^2}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

Goals

- Reduce randomization
- Maintain (improve) error bounds

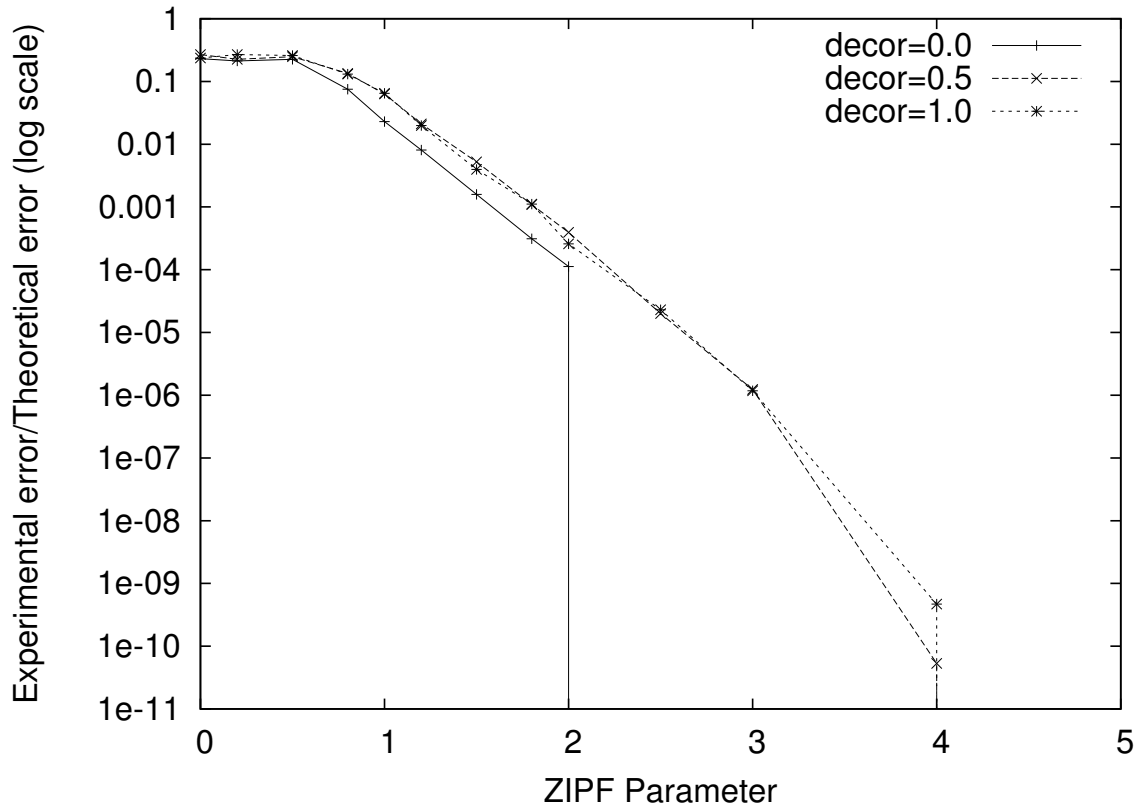
Experimental Comparison

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



Contributions

- Statistical analysis of sketch estimators
 - Probability distribution
- Derive tighter confidence bounds
 - Distribution-dependent
- Extensive experimental study
 - Synthetic & real data-sets

Confidence Bounds

Problem

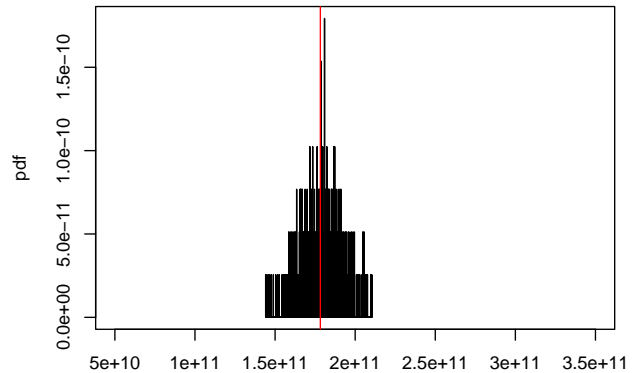
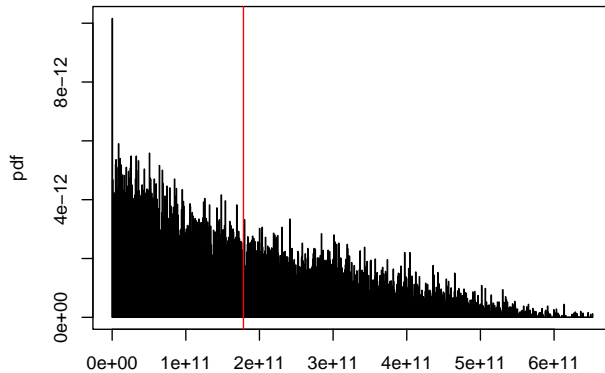
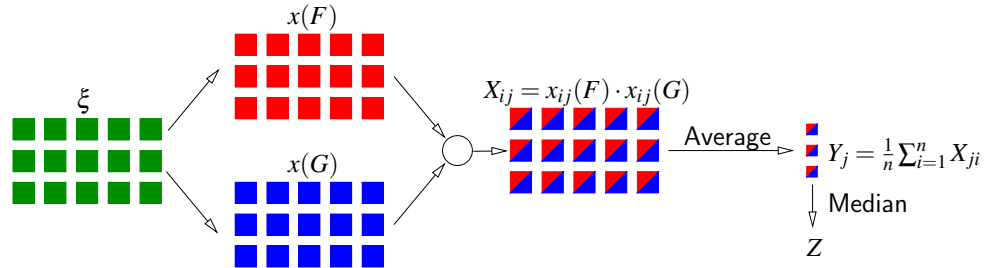
- X_1, X_2, \dots, X_n independent samples from random variable (estimator) X
- What is the true value (or expected value $E[X]$) of the estimator ?

Confidence bounds (intervals)

- true value $\in (f(X) \pm \varepsilon g(X))$ with probability $1 - \delta$, $\varepsilon, \delta \in (0, 1)$
- Frequency moments of X (expectation $E[X]$, variance $\text{Var}[X]$)
- Distribution-independent
 - Measure theory (Markov, Chebyshev, Chernoff, ...)
- Distribution-dependent
 - Statistics & sampling (cdf of the asymptotic distribution for X)

AGMS Sketches (AMS96,AGMS99)

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



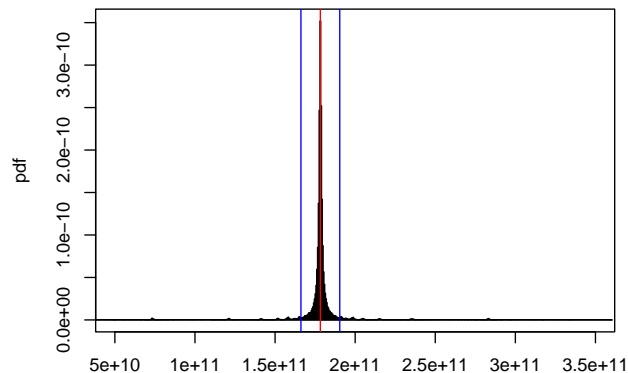
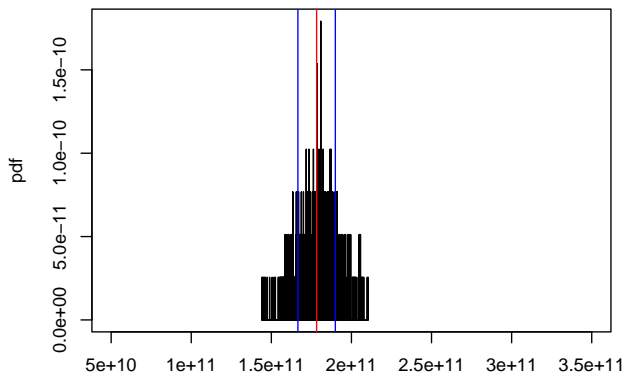
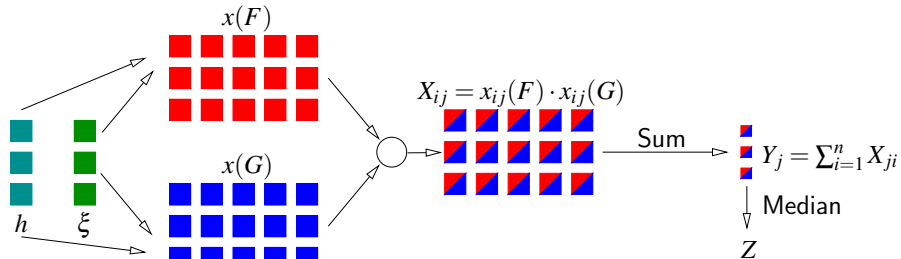
AGMS Sketches (AMS96,AGMS99)

Confidence bounds

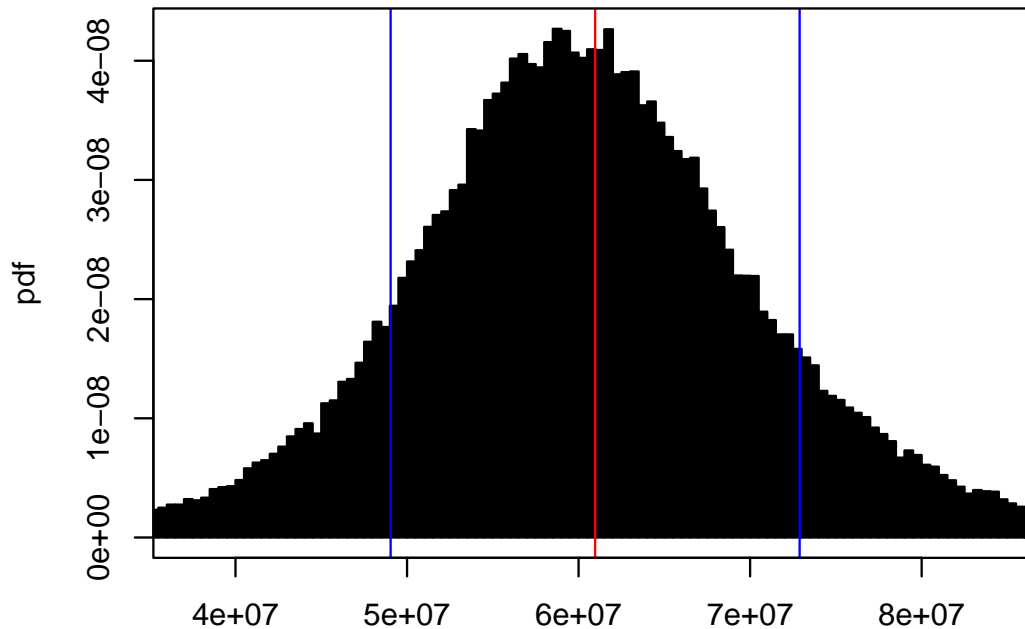
- Frequency moments of X : $E[X]$, $\text{Var}[X]$
- Estimator for true value: median of means Z
- Distribution-independent
 - Chebyshev & Chernoff bounds
 - $Z \in (\bar{f} \odot \bar{g} \pm \varepsilon \|\bar{f}\|_2 \|\bar{g}\|_2)$ with probability at least $1 - \delta$
- Distribution-dependent
 - Mean Central Limit Theorem (CLT), Median CLT, efficiency
 - Median confidence bounds are larger by $\sqrt{\frac{\pi}{2}} \approx 1.25$ for normal distribution
 - Practical estimator for $E[X]$: **mean** $Z = \frac{1}{n} \sum_{i=1}^n X_i$
- 95% distribution-dependent confidence bound is tighter by a factor of approx. 4

Fast-AGMS Sketches (CG05)

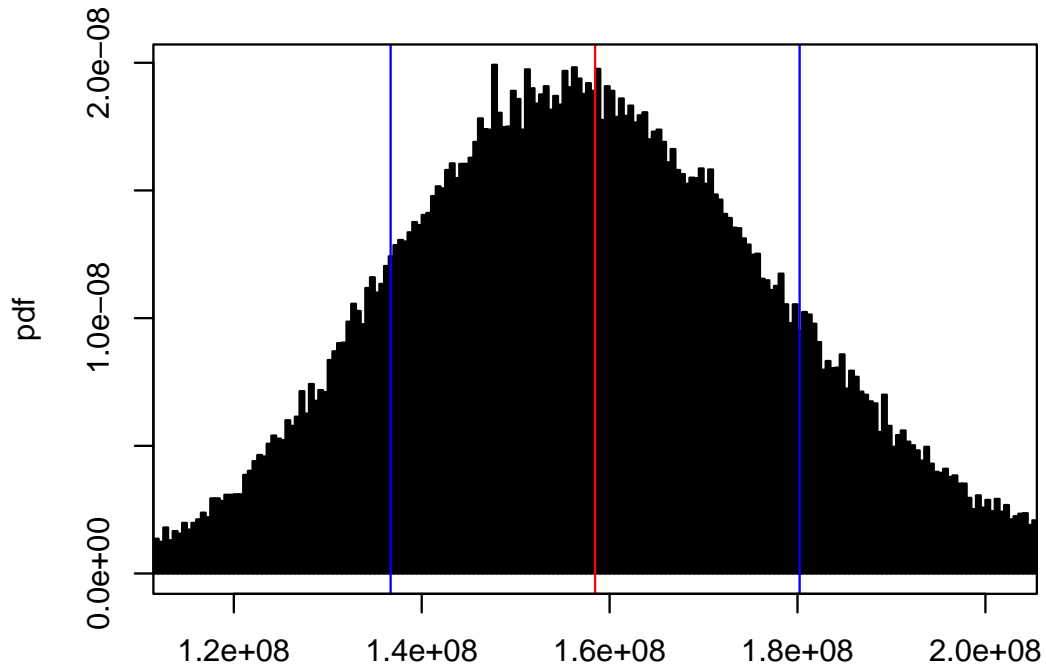
Setup: self-join size, Zipf=1.5, true result $\approx 1.8 \cdot 10^{11}$



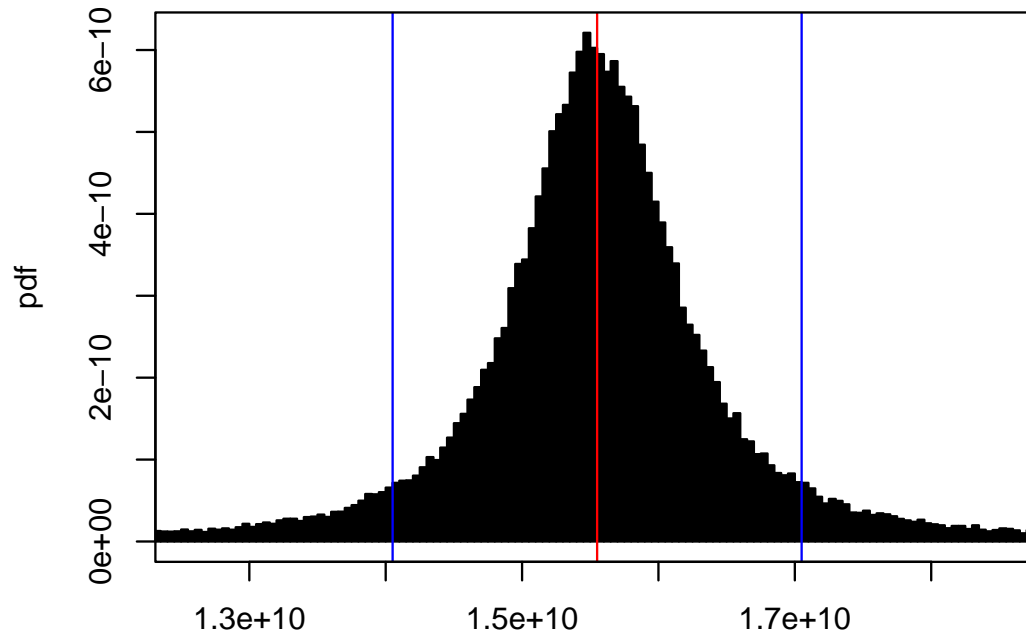
Zipf=0.0



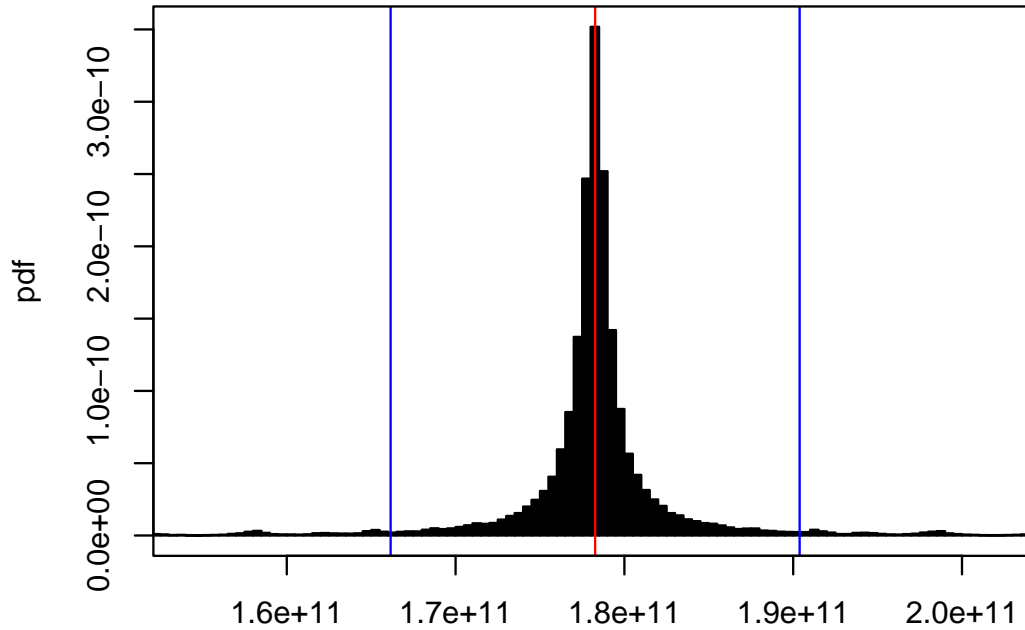
Zipf=0.5



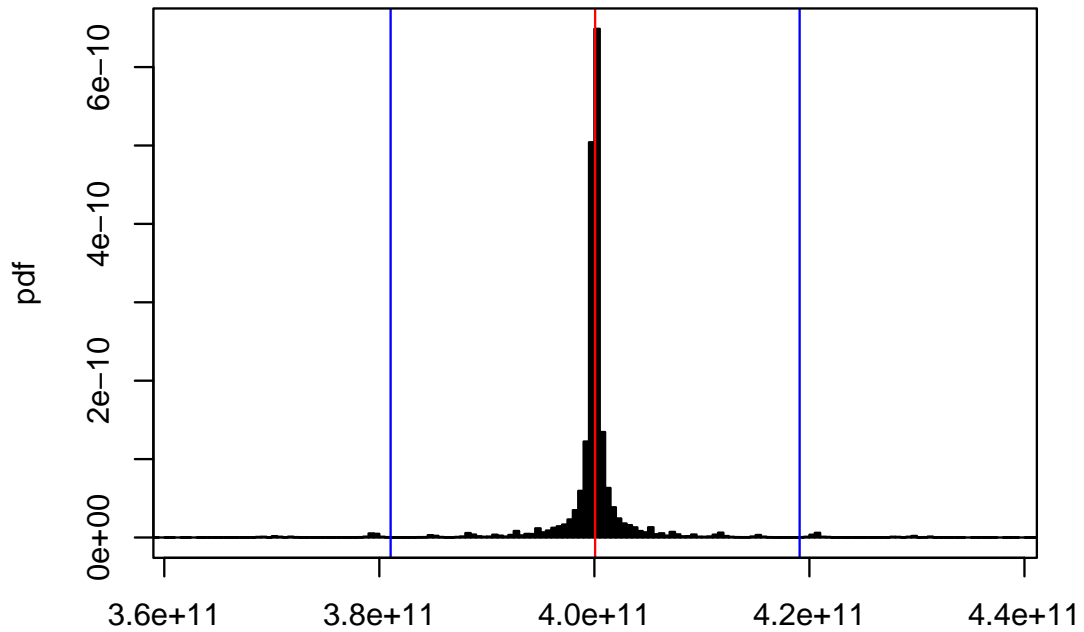
Zipf=1.0



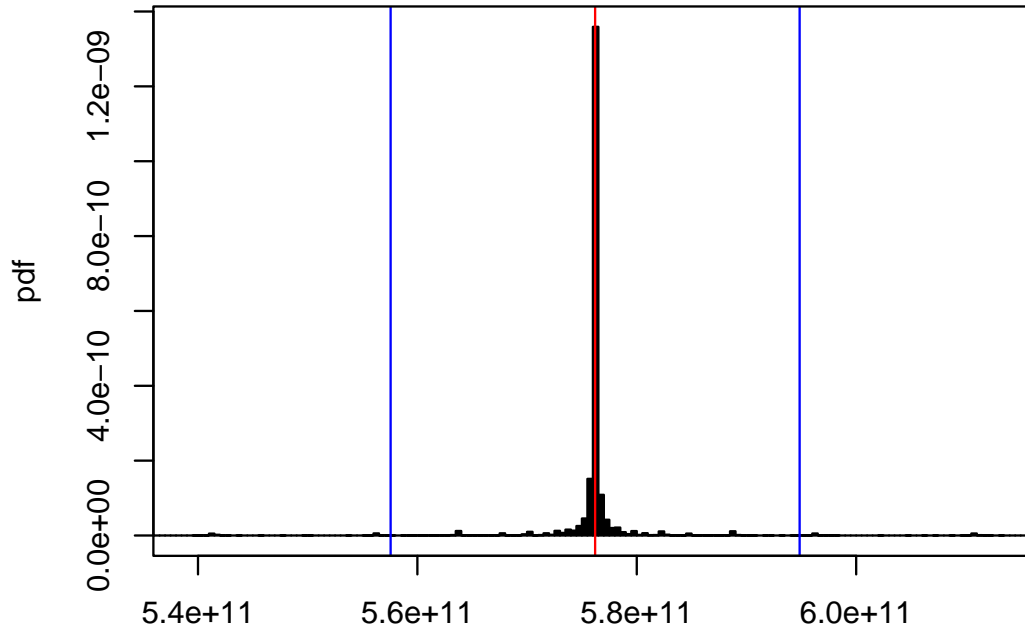
Zipf=1.5



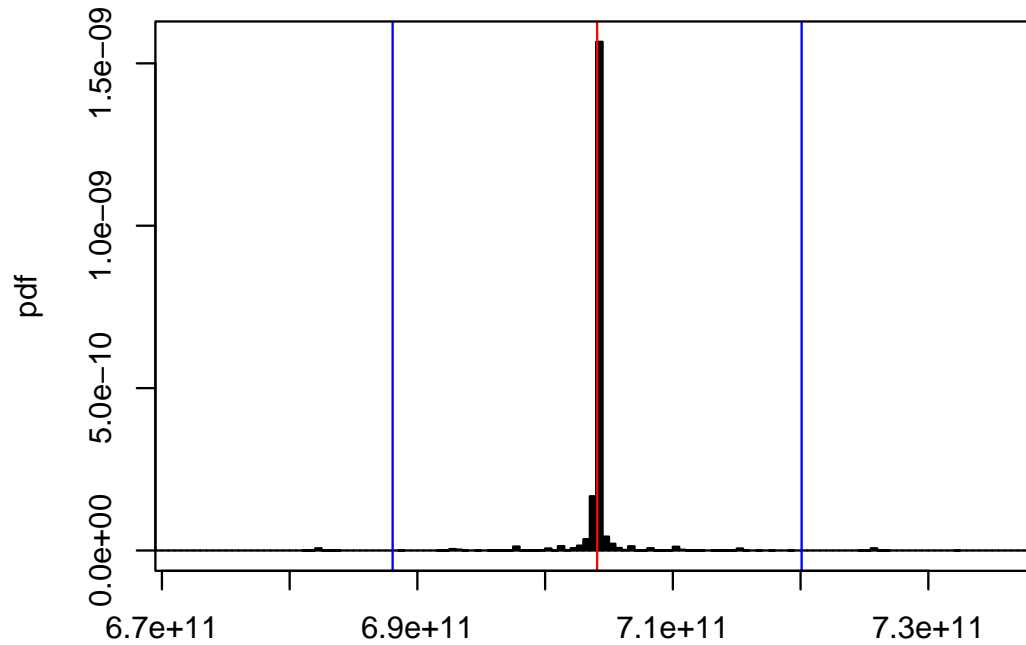
Zipf=2.0



Zipf=2.5



Zipf=3.0



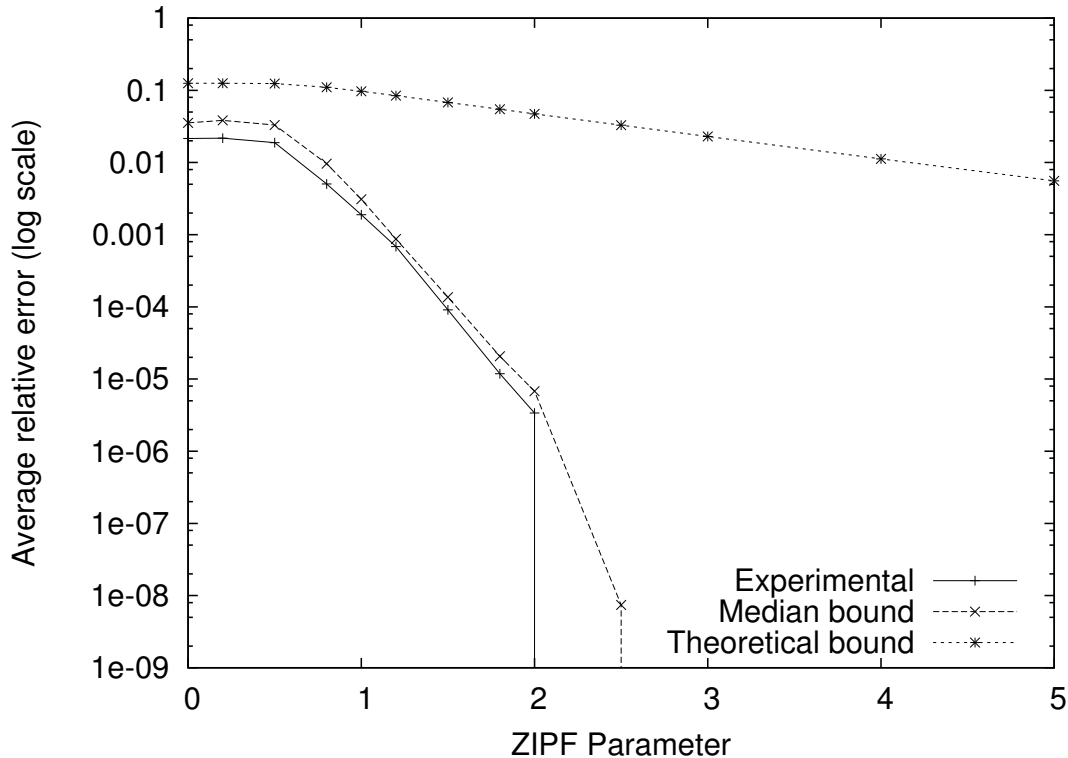
Fast-AGMS Sketches (CG05)

Confidence bounds

- Frequency moments of X : $E[X]$, $\text{Var}[X]$
- Estimator for true value: median Z
- Distribution-independent
 - Chebyshev & Chernoff bounds
 - $Z \in (\bar{f} \odot \bar{g} \pm \varepsilon \|\bar{f}\|_2 \|\bar{g}\|_2)$ 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



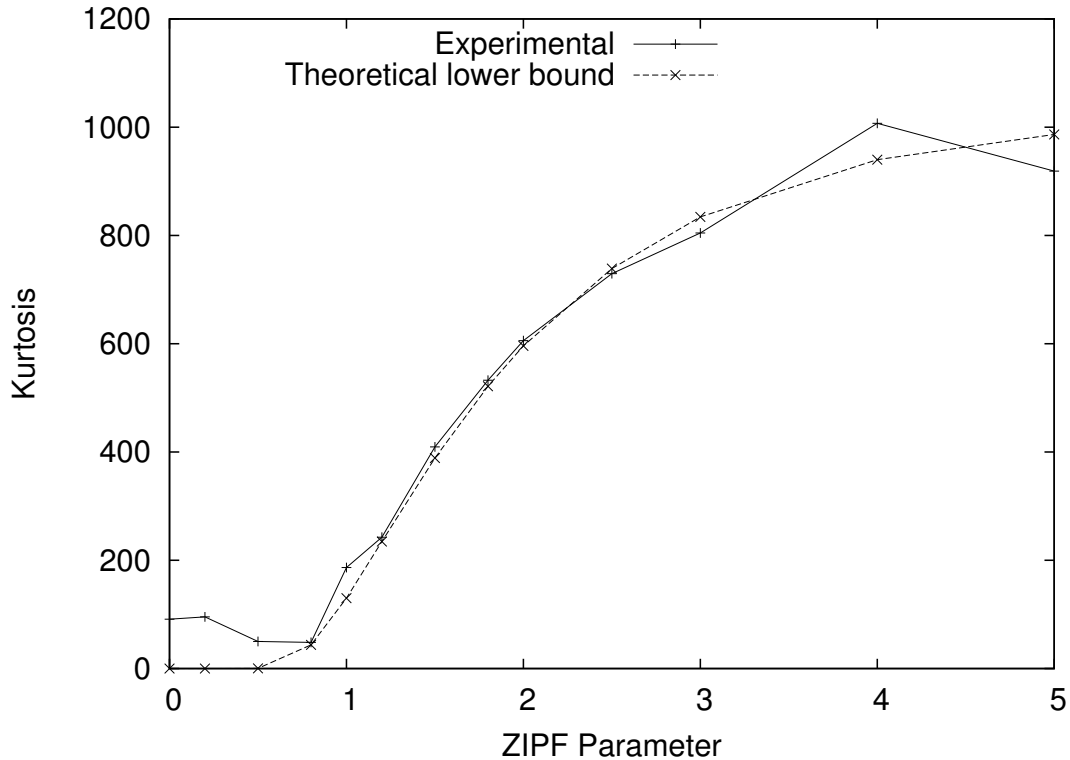
AGMS vs Fast-AGMS

Mean vs Median

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

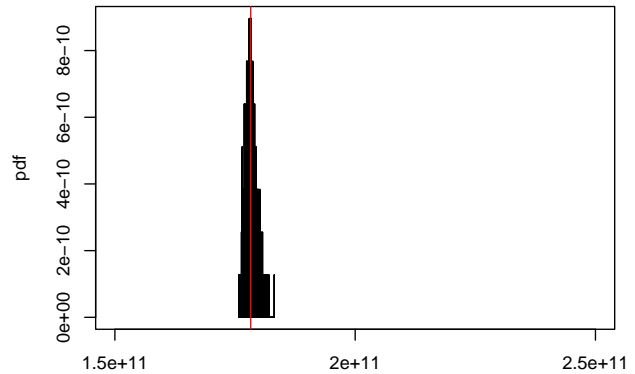
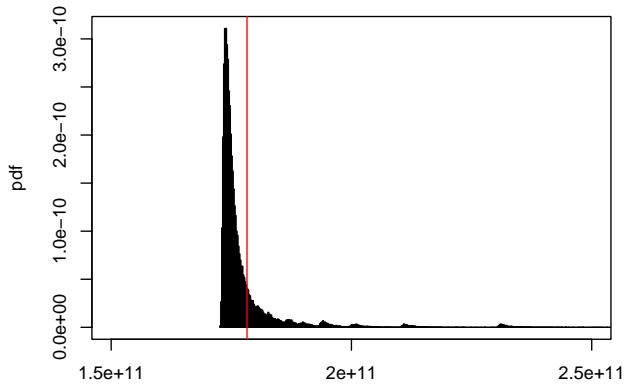
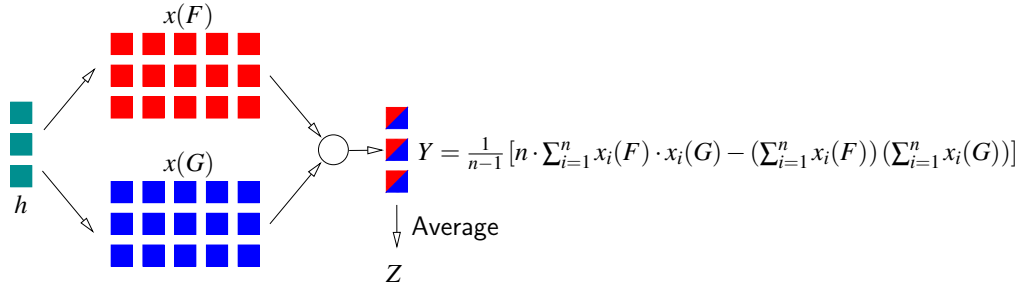
Fast-AGMS Kurtosis

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



Fast-Count Sketches (TZ04)

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



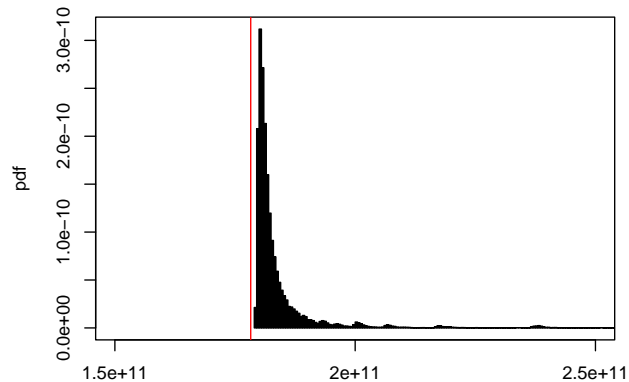
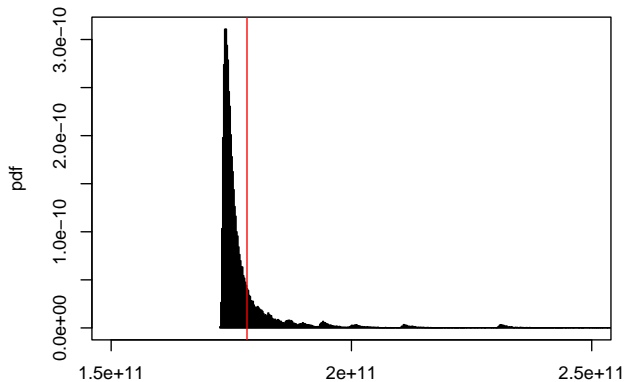
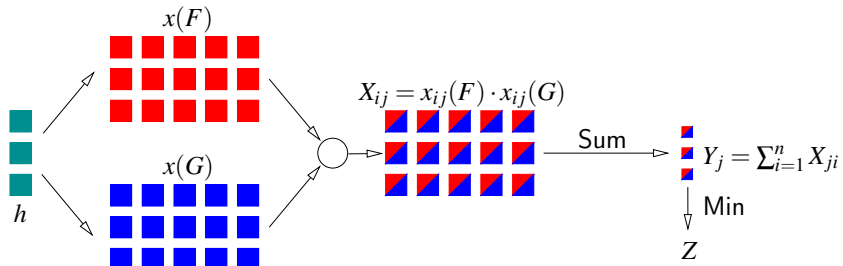
Fast-Count Sketches (TZ04)

Confidence bounds

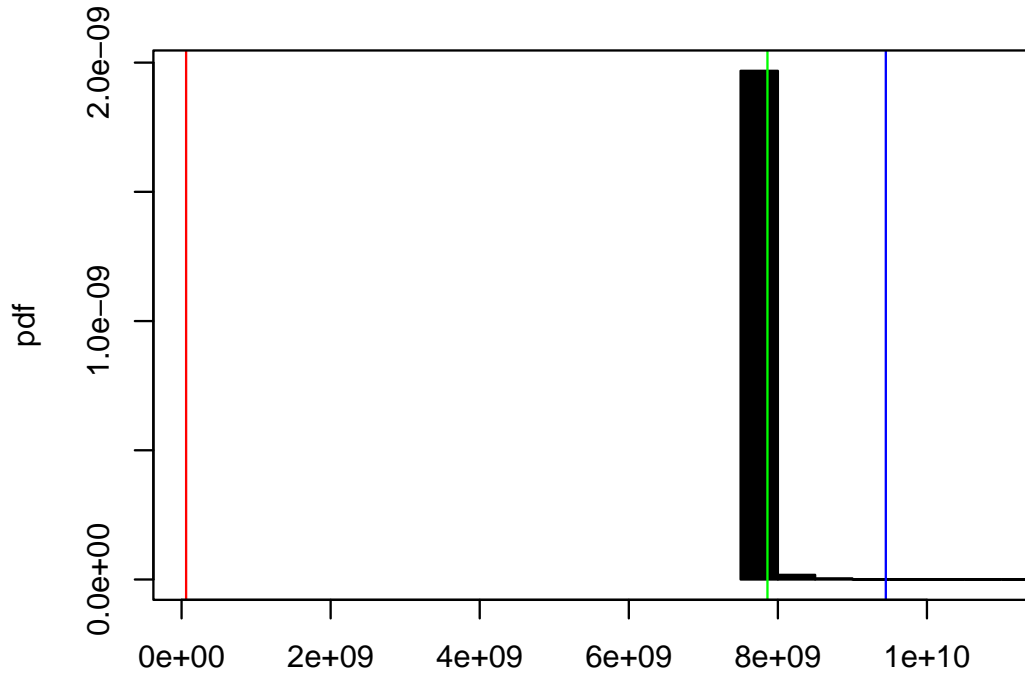
- Frequency moments of X : $E[X]$, $\text{Var}[X]$
- Estimator for true value: mean Z
- Distribution-independent
 - Chebyshev & Chernoff bounds
 - $Z \in (\bar{f} \odot \bar{g} \pm \varepsilon \|\bar{f}\|_2 \|\bar{g}\|_2)$ with probability at least $1 - \delta$
 - Identical to AGMS and Fast-AGMS sketches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

Count-Min Sketches (CM05)

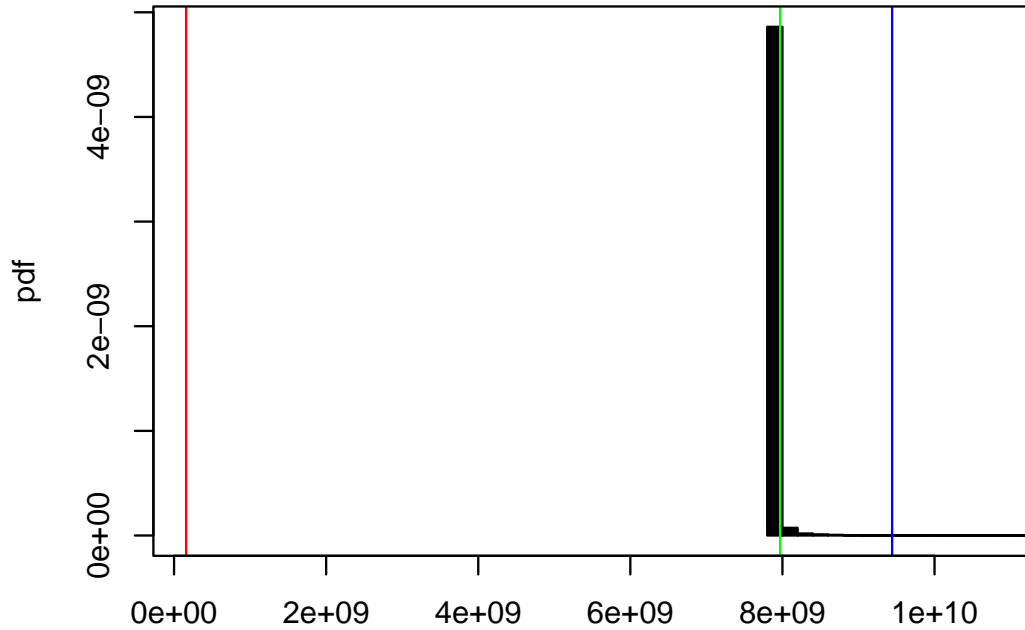
Setup: self-join size, Zipf=1.5, true result $\approx 1.8 \cdot 10^{11}$



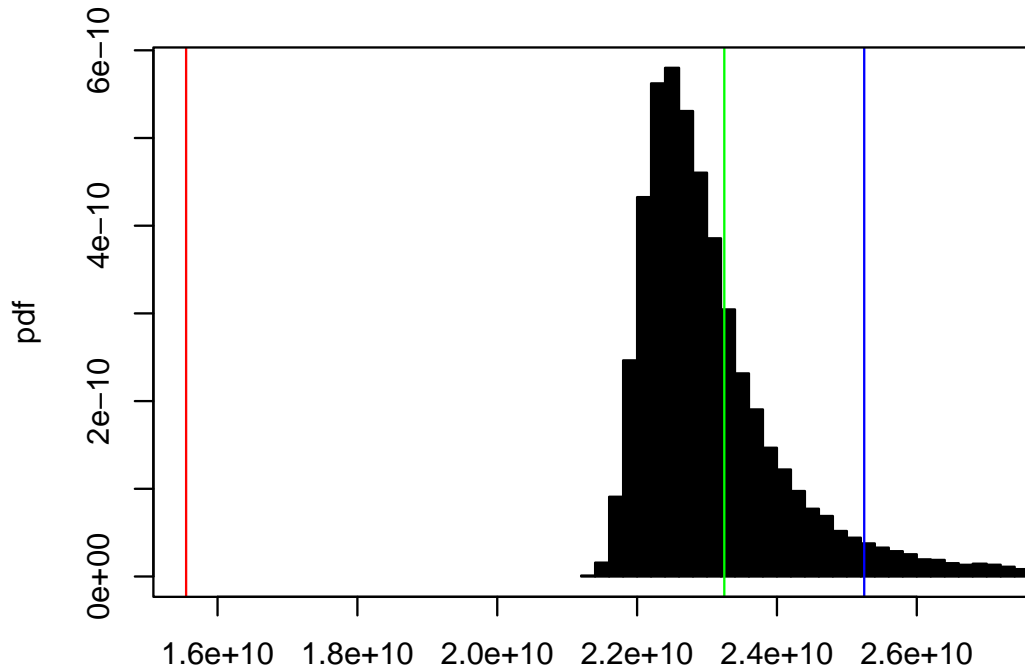
Zipf=0.0



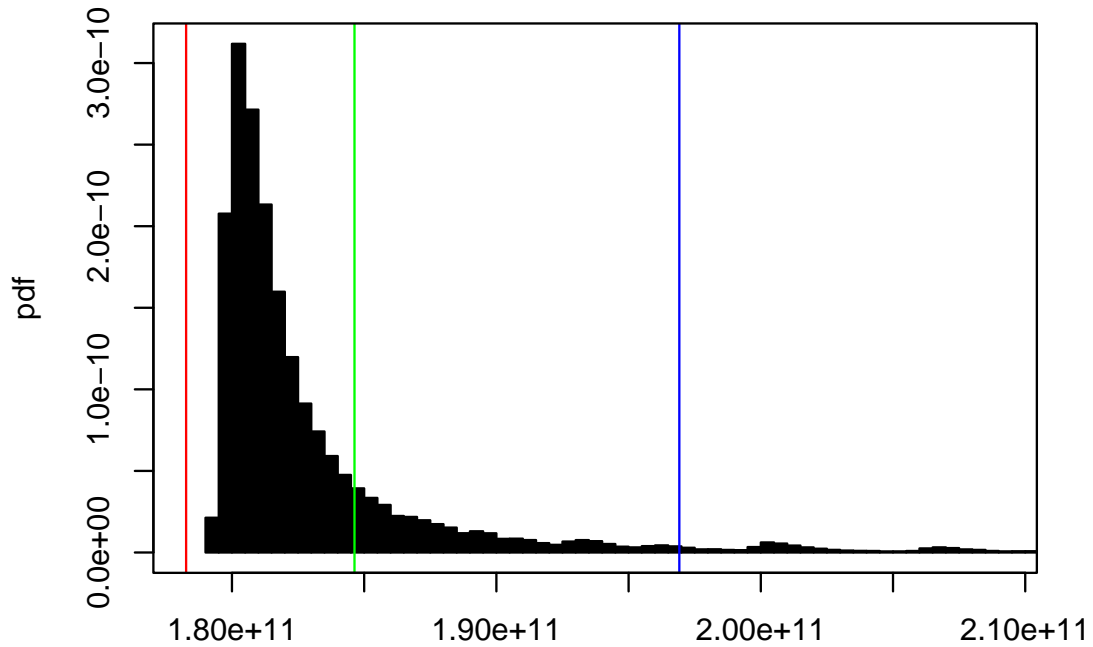
Zipf=0.5



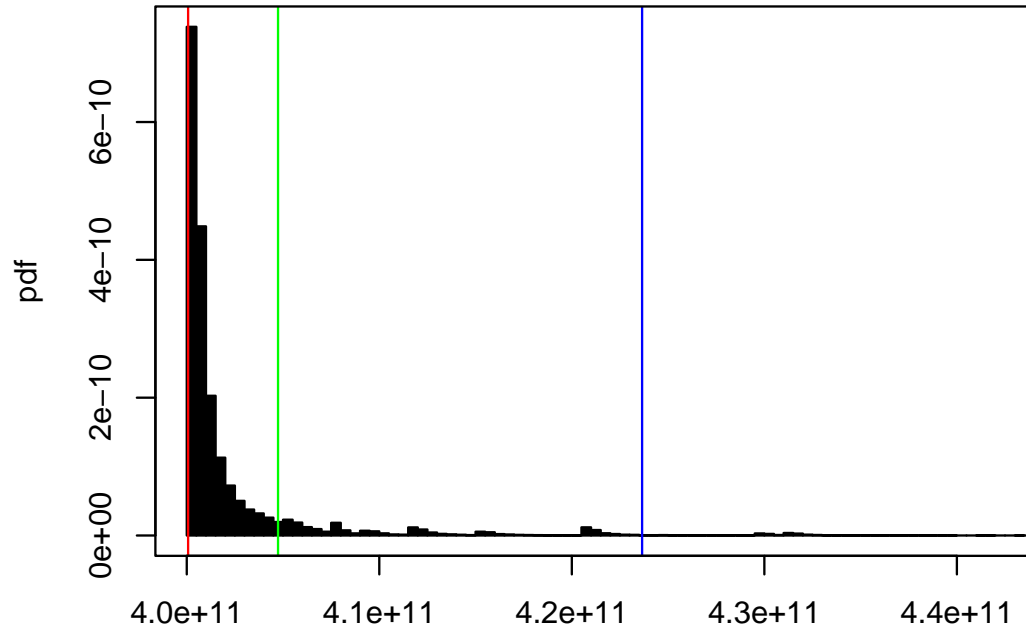
Zipf=1.0



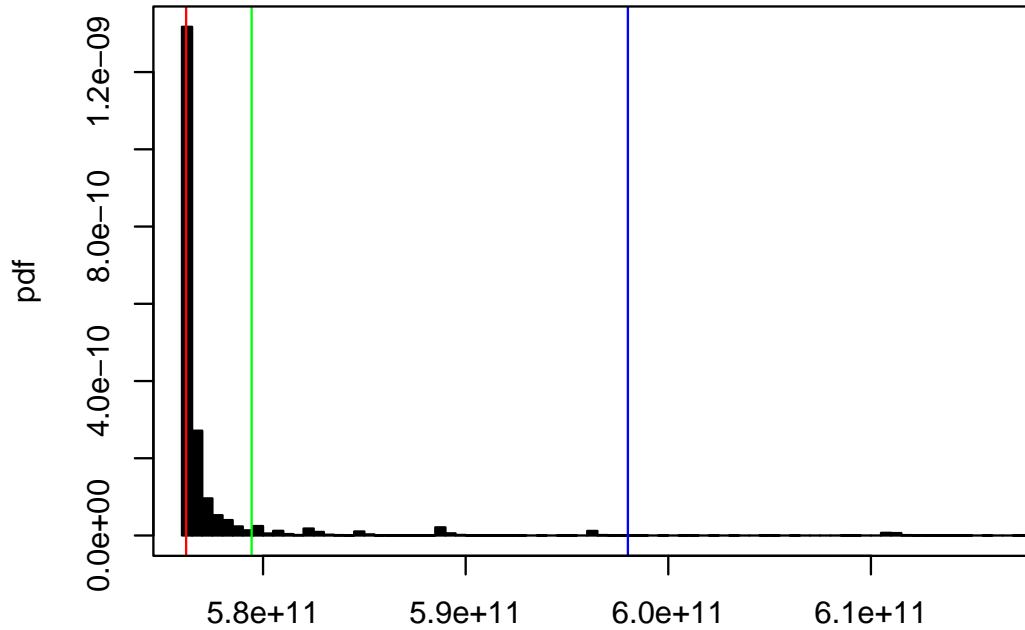
Zipf=1.5



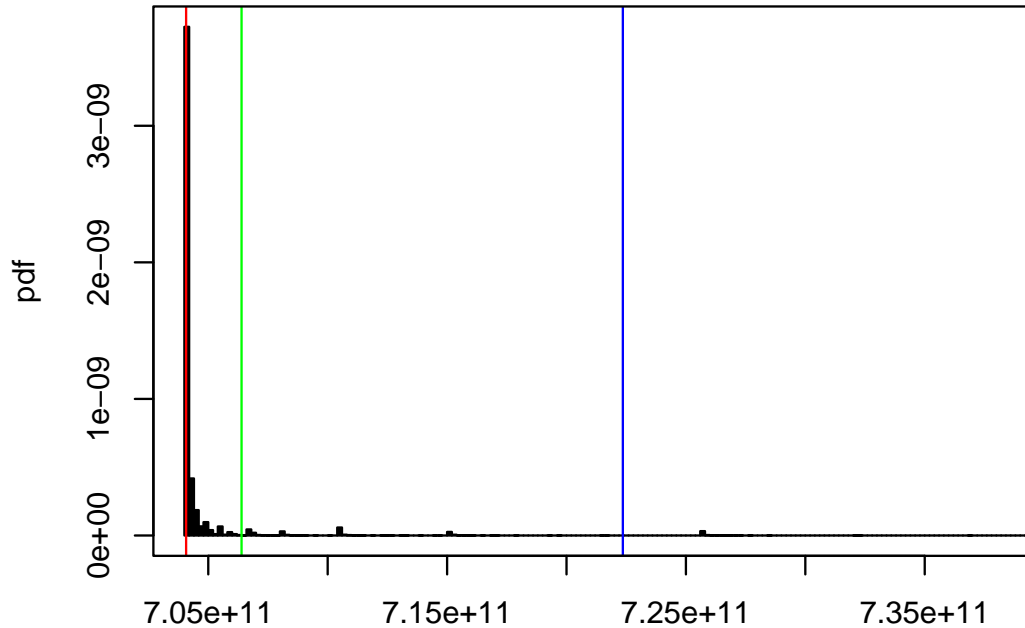
Zipf=2.0



Zipf=2.5



Zipf=3.0



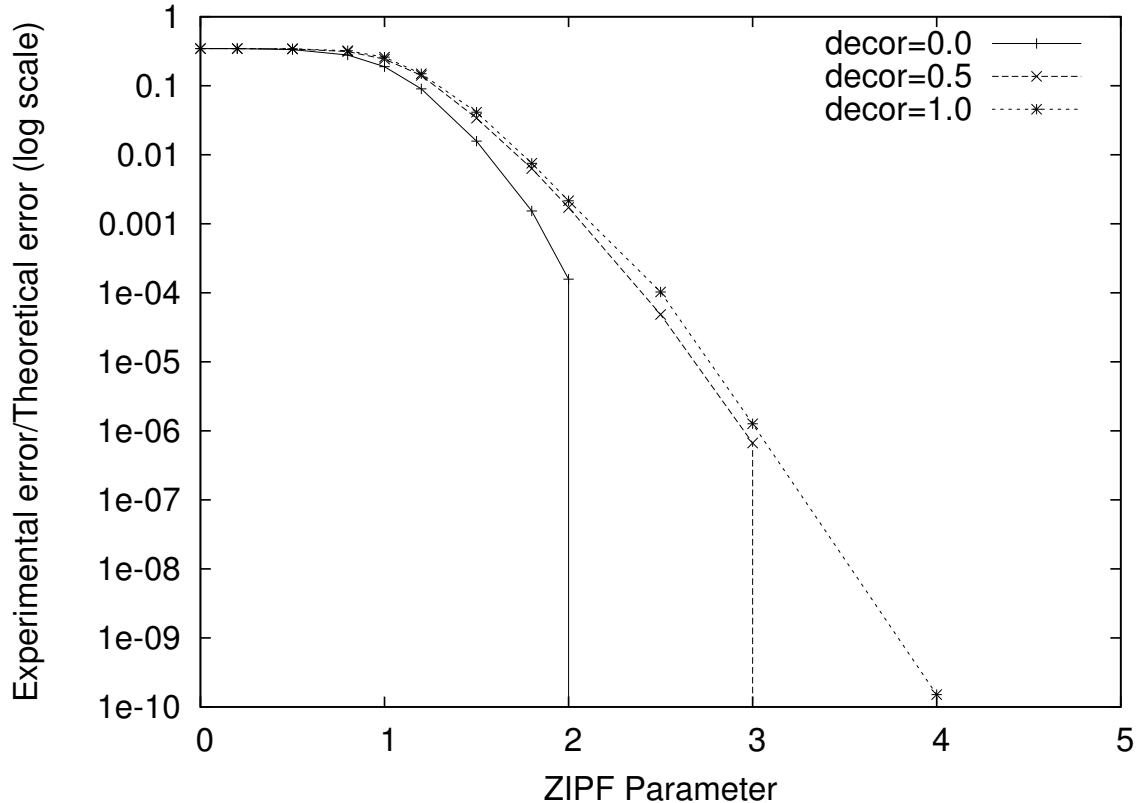
Count-Min Sketches (CM05)

Confidence bounds

- Frequency moments of X : $E[X]$, $\text{Var}[X]$
- Estimator for true value: minimum Z (Z is an **over-estimate**)
- Distribution-independent
 - Markov bounds
 - $\bar{f} \odot \bar{g} \leq Z \leq \bar{f} \odot \bar{g} + \varepsilon \|\bar{f}\|_1 \|\bar{g}\|_1$ with probability at least $1 - \delta$
 - Variance is identical to AGMS and Fast-AGMS sketches
- Distribution-dependent
 - Minimum CLT [C01]
 - Asymptotic distribution of Z is Generalized Extreme Value (GEV)

Count-Min Sketches (CM05)

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



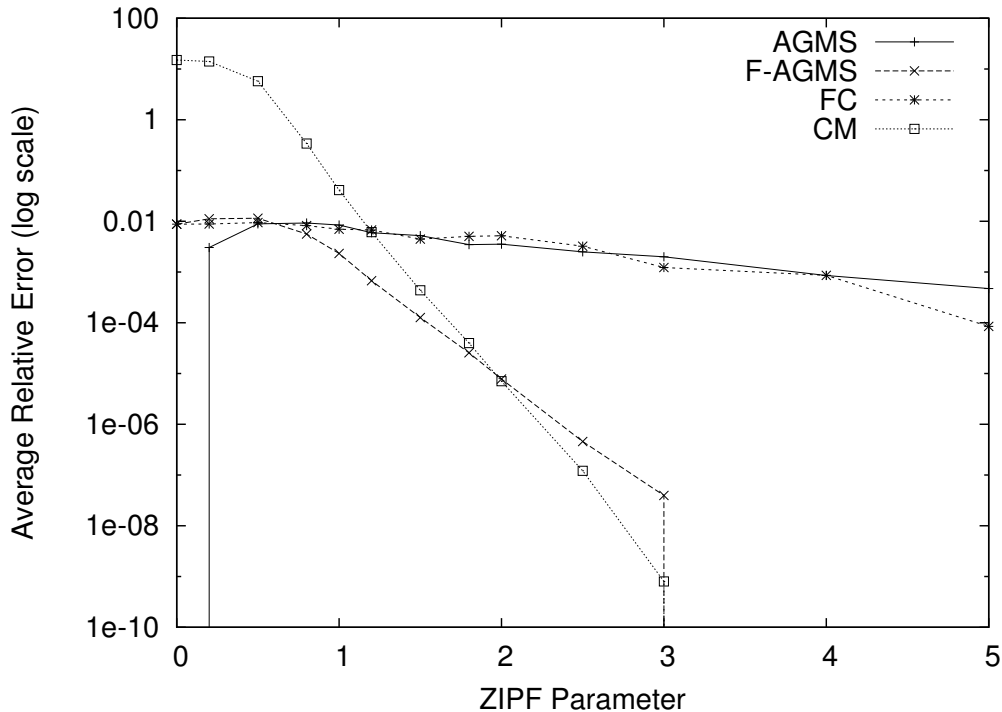
Experimental Study

First extensive empirical study of sketches

- Absolute comparison – relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)
 - Data skew (Zipf coefficient)
 - Data correlation
 - Memory budget
 - Update time
- Relative comparison
- Data sets
 - Synthetic (Zipf, correlation)
 - Real

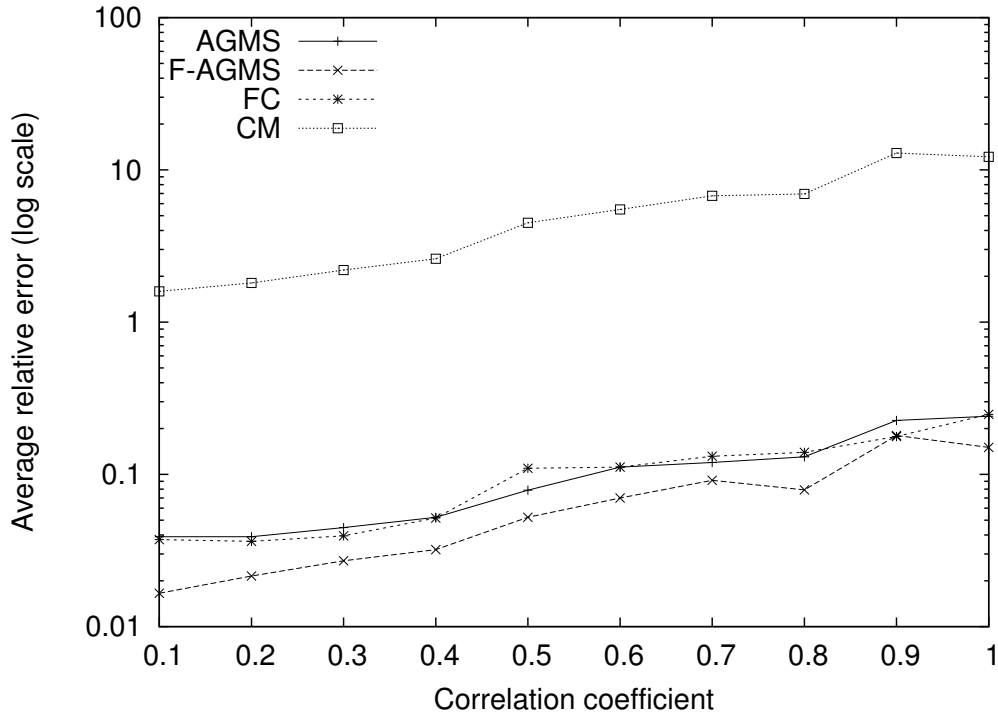
Data Skew

Setup: self-join size, sketch size (row=21, col=1024), relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)



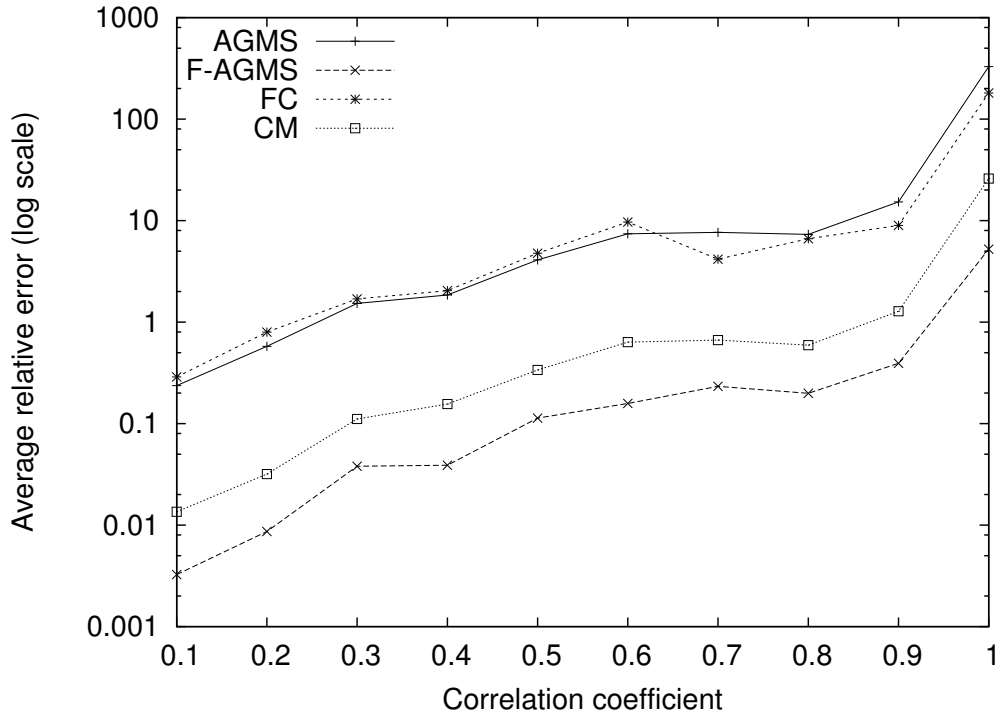
Data Correlation Zipf = 0.8

Setup: size of join, sketch size (row=21, col=1024), relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)



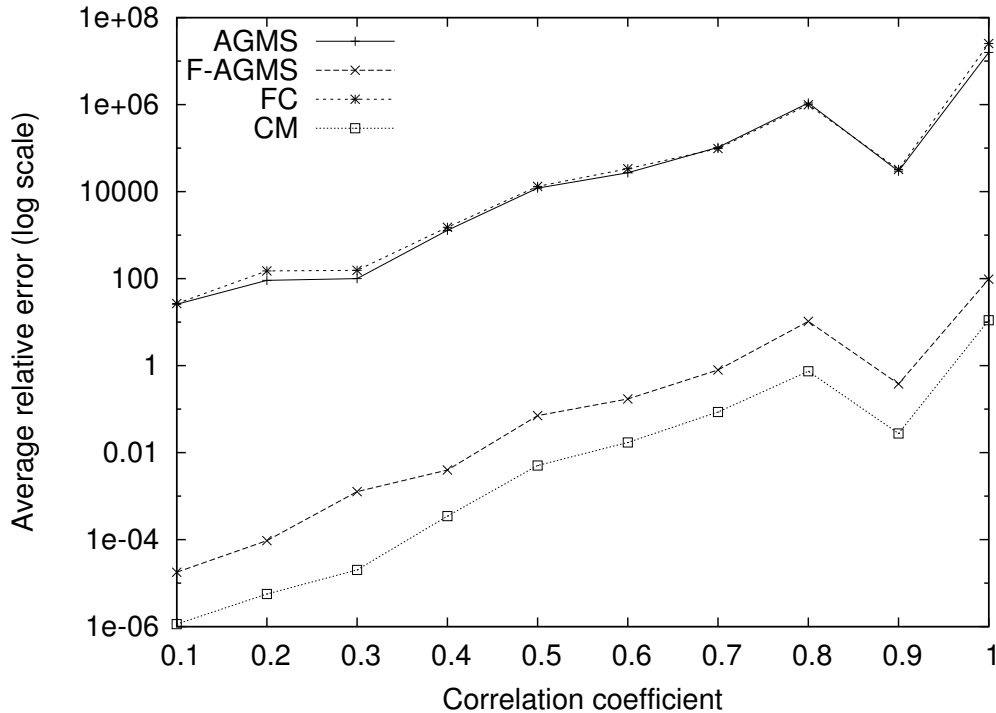
Data Correlation Zipf = 1.5

Setup: size of join, sketch size (row=21, col=1024), relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)



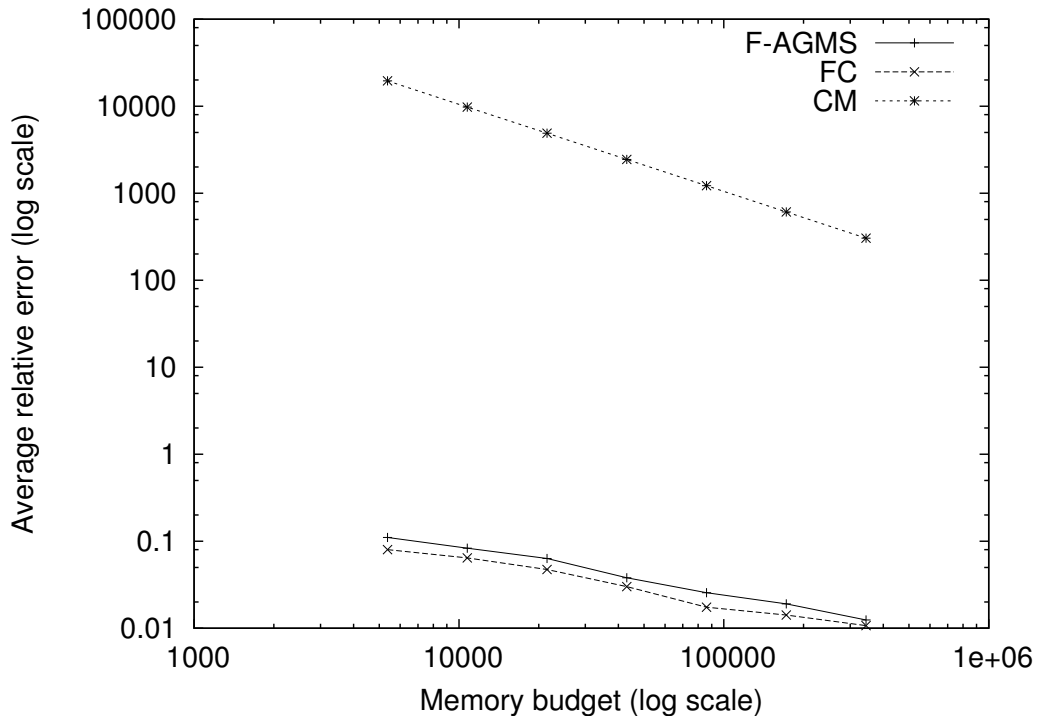
Data Correlation Zipf = 3.0

Setup: size of join, sketch size (row=21, col=1024), relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)



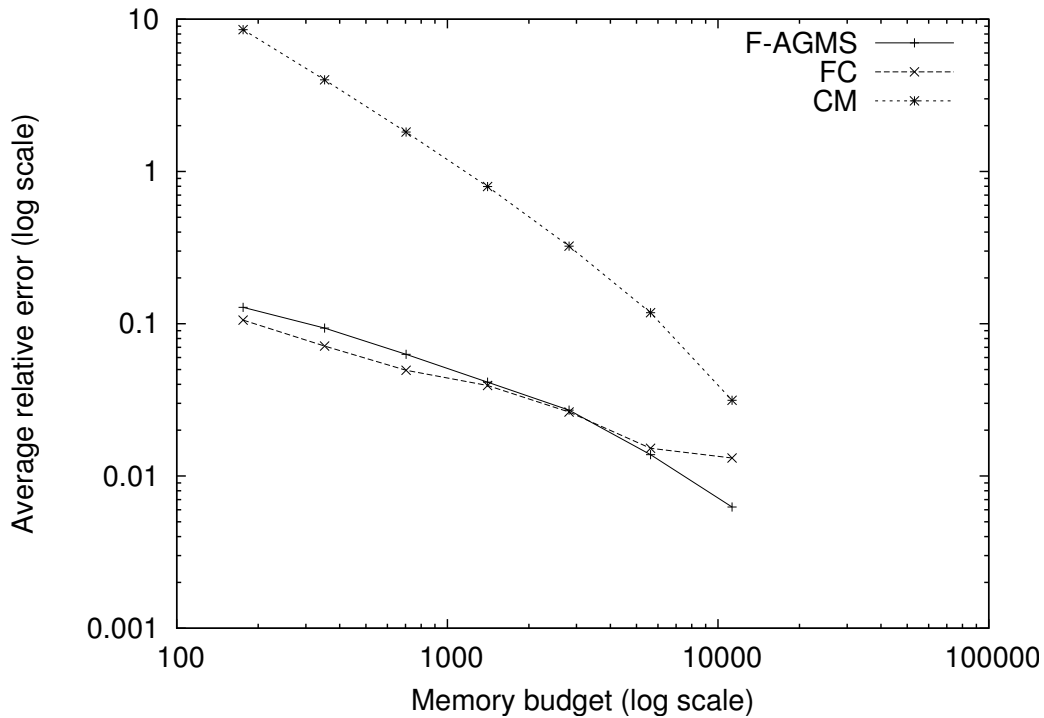
Memory Budget ((EN06) memory unpeaked)

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



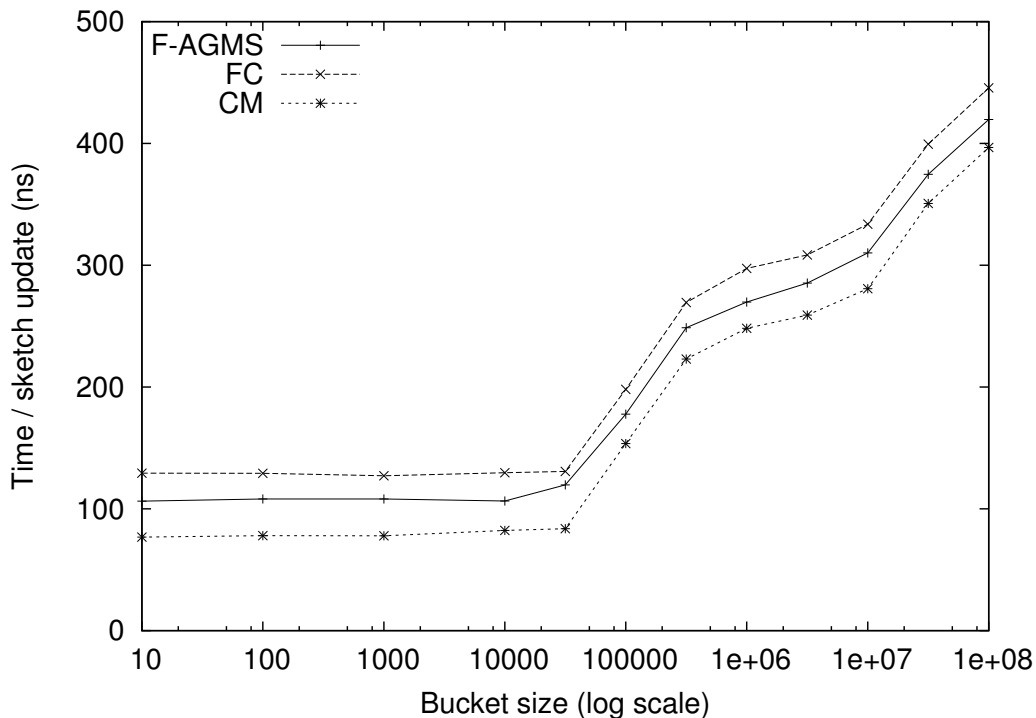
Memory Budget (census data)

Setup: size of join, sketch size (row=11), relative error ($\frac{|\text{actual}-\text{estimate}|}{\text{actual}}$)



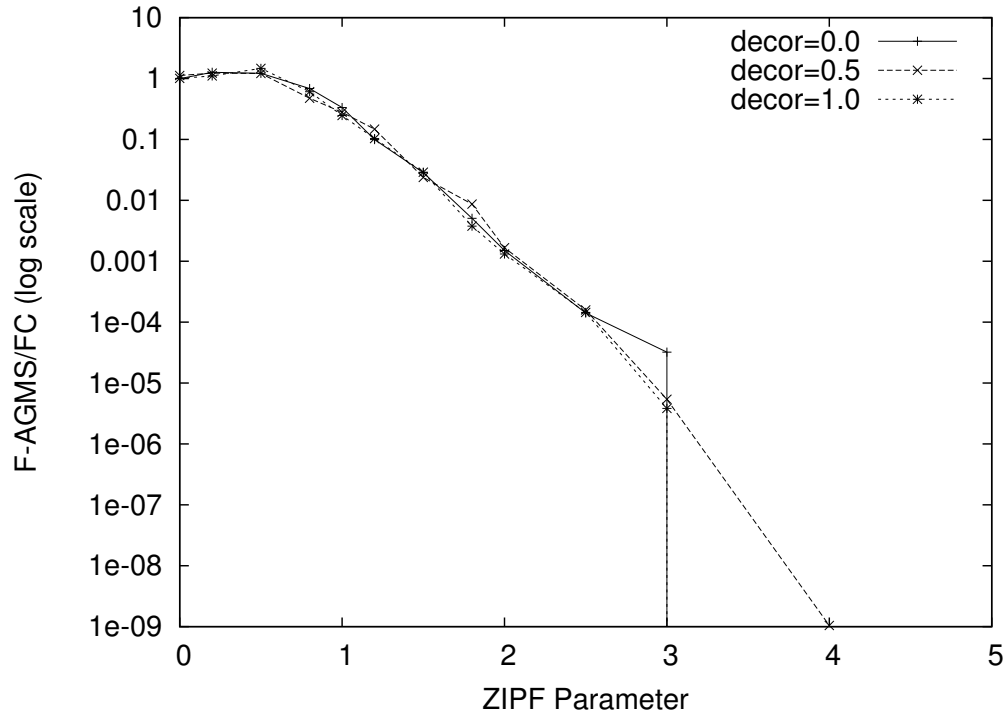
Update Time

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



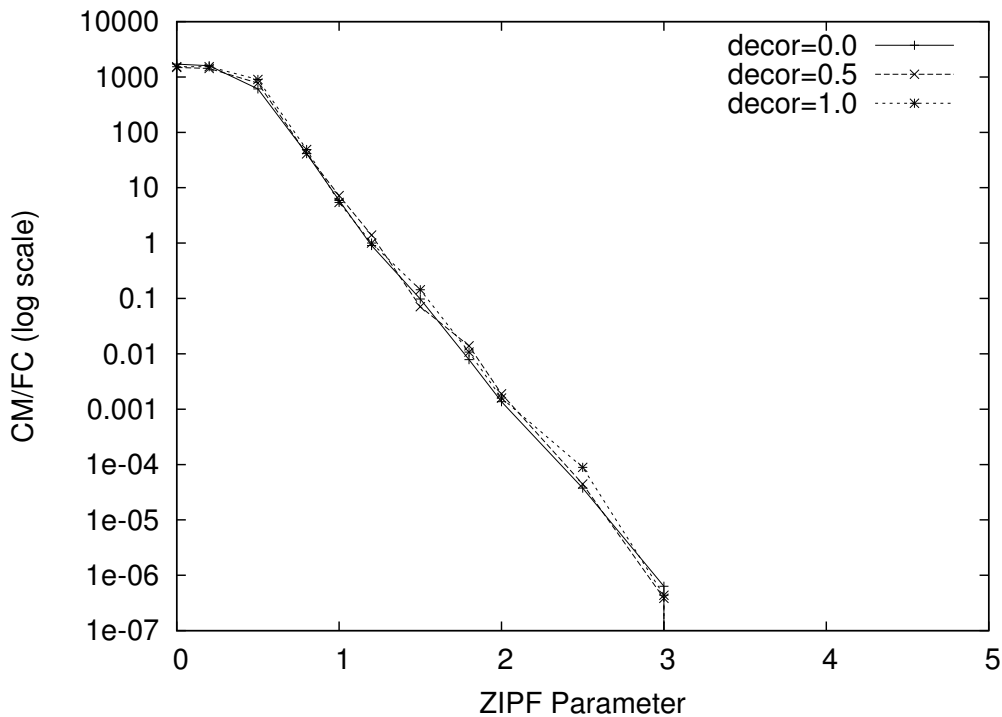
F-AGMS vs FC (AGMS)

Setup: size of join, sketch size (row=21, col=1024)



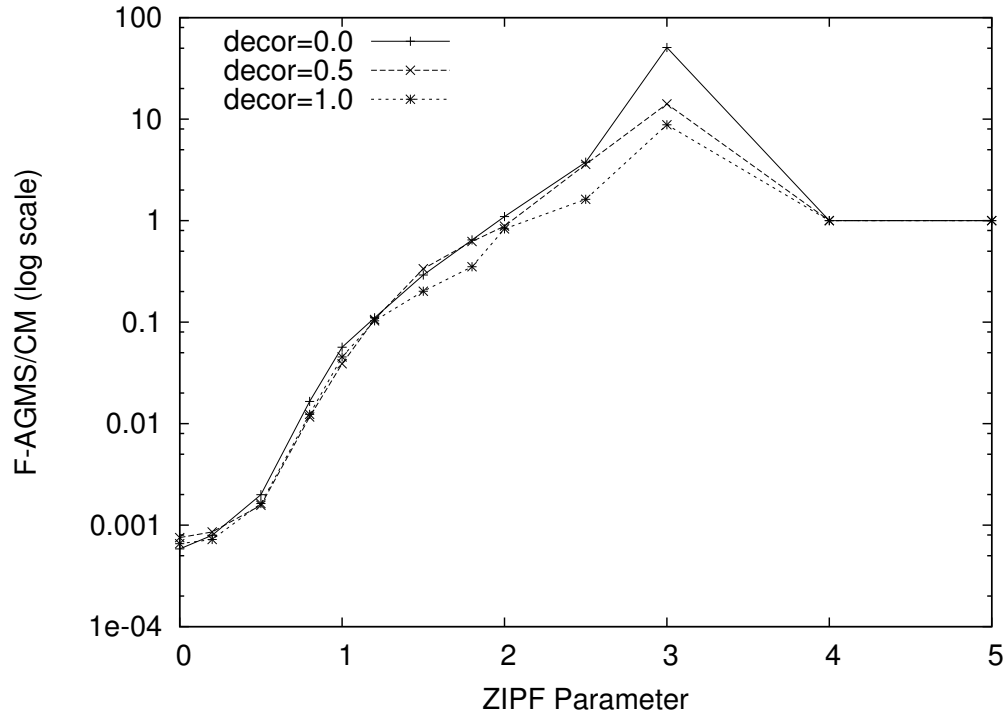
CM vs FC (AGMS)

Setup: size of join, sketch size (row=21, col=1024)



F-AGMS vs CM

Setup: size of join, sketch size (row=21, col=1024)



Experimental Study

Findings

- AGMS and FC have similar performance (FC has better update time)
- CM is the best for high skew data and the worst for low skew data
- F-AGMS is always (close to) the best
- Sketches are performant for correlated data
 - Skew is the main factor affecting performance
 - Correlation has the same effect
- Update time is not an issue

Conclusions

Statistical analysis

- Estimator distribution
- Distribution-dependent confidence bounds (where possible)

Extensive experimental study

- Comparison of sketch estimators

Recommendation

- Use F-AGMS sketches

Questions