Two-Level Sampling for Join Size Estimation

Yu Chen, Ke Yi

Hong Kong University of Science and Technology

Introduction

Problem

- PK-FK join

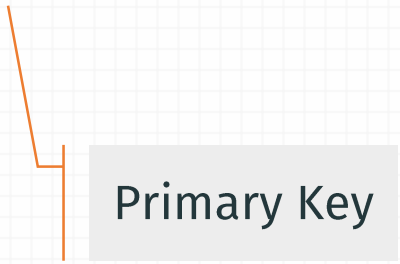
$$\text{Customer} \bowtie \text{Order}$$

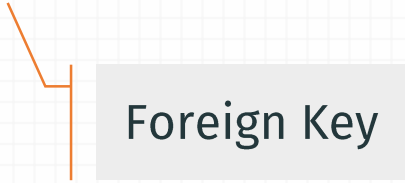
Customer

Custkey	Name	Phone	Age
1	Lizabeth	(143) 254-5453	41
2	Elliott	(422) 809-7954	65
3	Helga	(906) 893-5224	20
4	Parker	(346) 386-9233	47
5	Wilford	(239) 710-1702	22

Order

Ordkey	Custkey	TotalPrice
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418



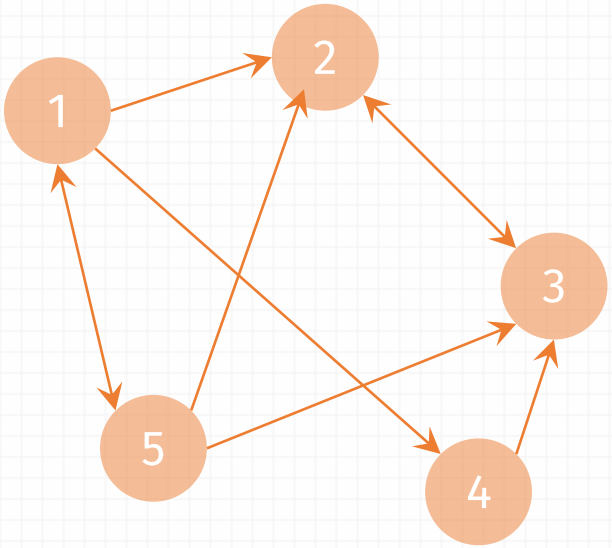


Problem

- Many-many join

Twitter

Follower_id	Followee_id
1	2
1	4
1	5
2	3
3	2
4	5
5	1
5	2
5	3

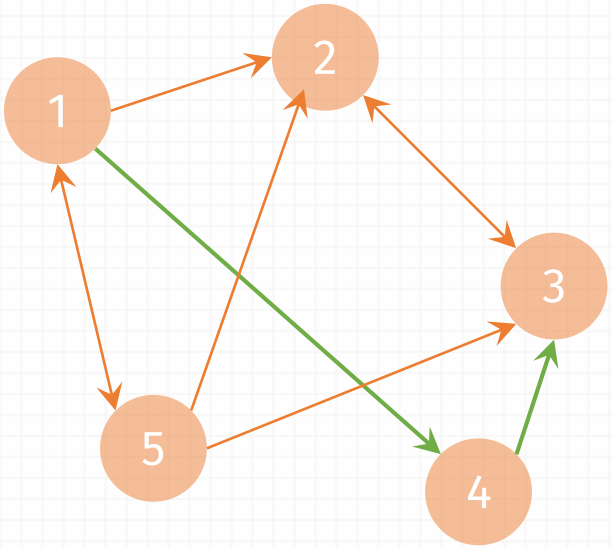


Problem

- Many-many join

Twitter

Follower_id	Followee_id
1	2
1	4
1	5
2	3
3	2
4	5
5	1
5	2
5	3



Problem

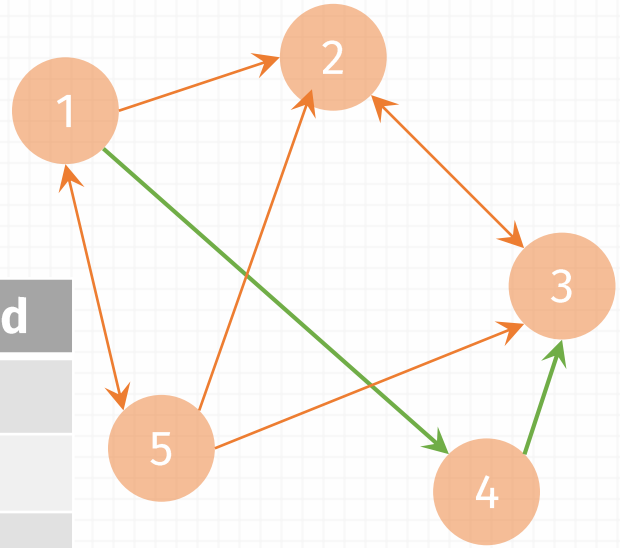
- Many-many join

T1

Follower_id	Followee_id
1	2
1	4
1	5
2	3
3	2
4	5
5	1
5	2
5	3

T2

Follower_id	Followee_id
1	2
1	4
1	5
2	3
3	2
4	5
5	1
5	2
5	3



$$T1 \bowtie_{\text{T1.Followee\_id} = \text{T2.Follower\_id}} T2$$

Problem

- Join with selection predicates

Customer

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22

Order

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418

Ord key	Cust key	Total Price	Name	...
3	5	420	Wilford	...

$$\text{Customer} \bowtie_{\text{Age}<=35 \text{ AND } \text{TotalPrice}>200} \text{Order}$$

Selection Predicates are given at query time

Problem

- Notations

A

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22

B

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418

$B(v)$

$|B(v)| = b_v$

$|B(2)| = b_2 = 2$

$\sigma_{c_A}(A) \bowtie \sigma_{c_B} B$

Ord key	Cust key	Total Price	Name	...
3	5	420	Wilford	...

Goal:
Estimate Join size at query time

$$J = |\sigma_{c_A}(A) \bowtie \sigma_{c_B} B| = \Sigma_v a_v^{c_A} b_v^{c_B}$$

Related Work

Main approaches

- Sketching
- Synopsis
- Random sampling

Main approaches

- Sketching
- Synopsis
- Random sampling

Sketching [Rusu et al. 2008]

- Build one sketch for each table on the join attribute
- Give very accurate estimates for join without predicates

Sketching

- Build one sketch for each table on the join attribute
- Dealing with selection predicates
  - Build sketch on ‘imaginary’ tables[Dobra et al. 2002]

Sketching

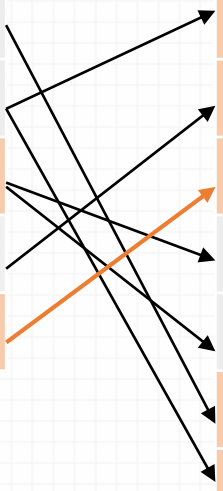
- Build one sketch for each table on the join attribute
- Dealing with selection predicates
  - Build sketch on ‘imaginary’ tables

$$\sigma_{\text{Age}<=35}(Customer) \bowtie \sigma_{\text{TotalPrice}>200}(Order)$$

Customer ⋈ Order

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418



Sketching

- Build one sketch for each table on the join attribute
- Dealing with selection predicates
  - Build sketch on ‘imaginary’ tables

$$\sigma_{\text{Age}<=35}(Customer) \bowtie \sigma_{\text{TotalPrice}>200}(Order)$$

Customer*

Age
1
2
…
21
20
…
34
35

$\bowtie$

Customer

Cust key	Name	…	Age
1	Lizabeth	…	41
2	Elliott	…	65
3	Helga	…	20
4	Parker	…	47
5	Wilford	…	22

$\bowtie$

Order

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418

$\bowtie$

Order*

Total Price
201
202
203
…
MAX_INT

Sketching

- Build one sketch for each table on the join attribute
- Dealing with selection predicates
  - Build sketch on ‘imaginary’ tables

$$\sigma_{\text{Age}<=35}(Customer) \bowtie \sigma_{\text{TotalPrice}>200}(Order)$$

Customer

Cust key	Name	…	Age
1	Lizabeth	…	41
2	Elliott	…	65
3	Helga	…	20
4	Parker	…	47
5	Wilford	…	22

$\bowtie$

Order

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418

2-table join

4-table join

Error grows rapidly

Sketching

- Build one sketch for each table on the join attribute

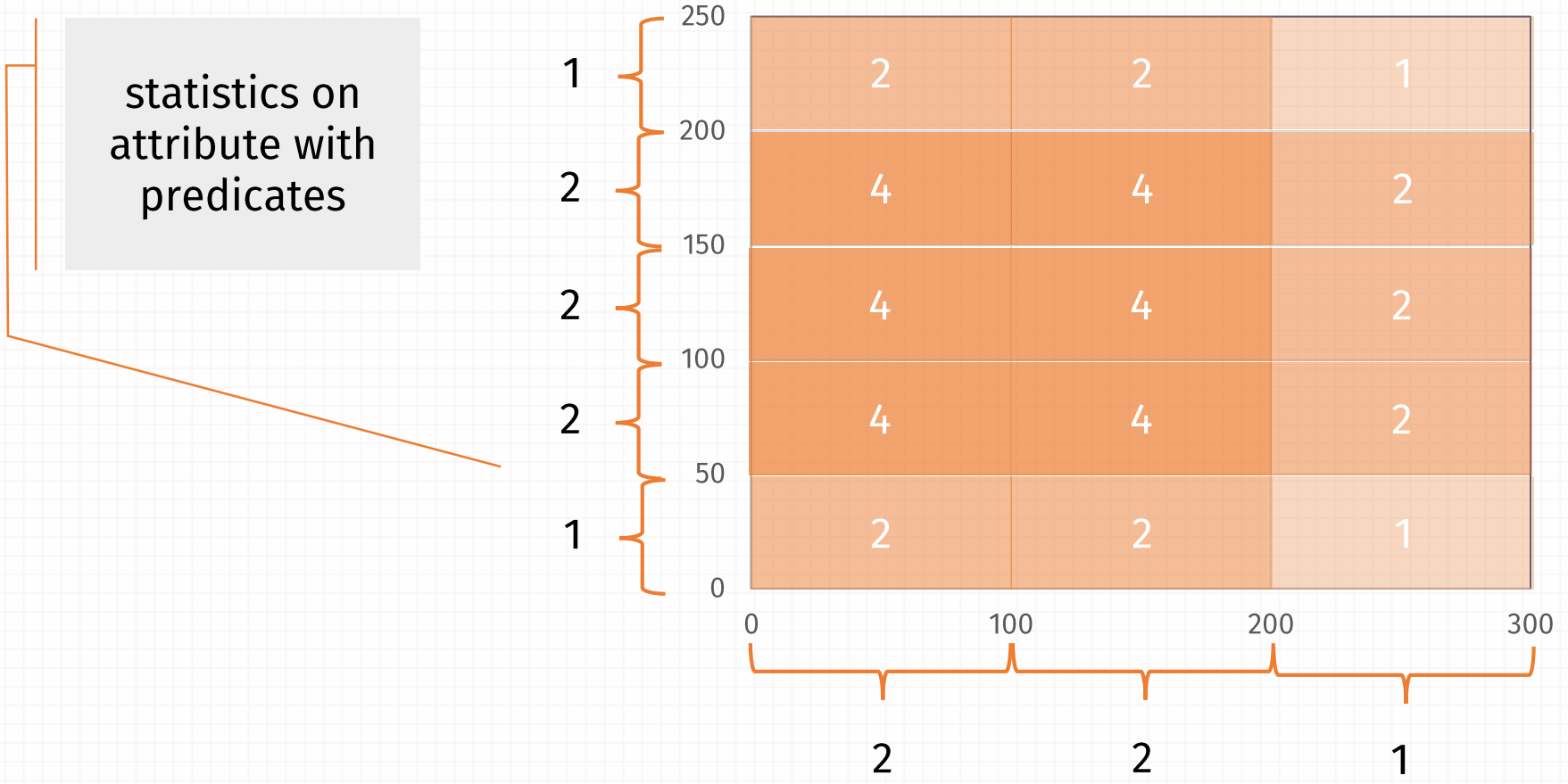
- Dealing with selection predicates
  - Build sketch on ‘imaginary’ tables at query time
  - Estimation error grows rapidly
  - Only handle equality/range predicate
    - “Comments LIKE %complain%”...

Main approaches

- Sketching
- Synopsis
- Random sampling

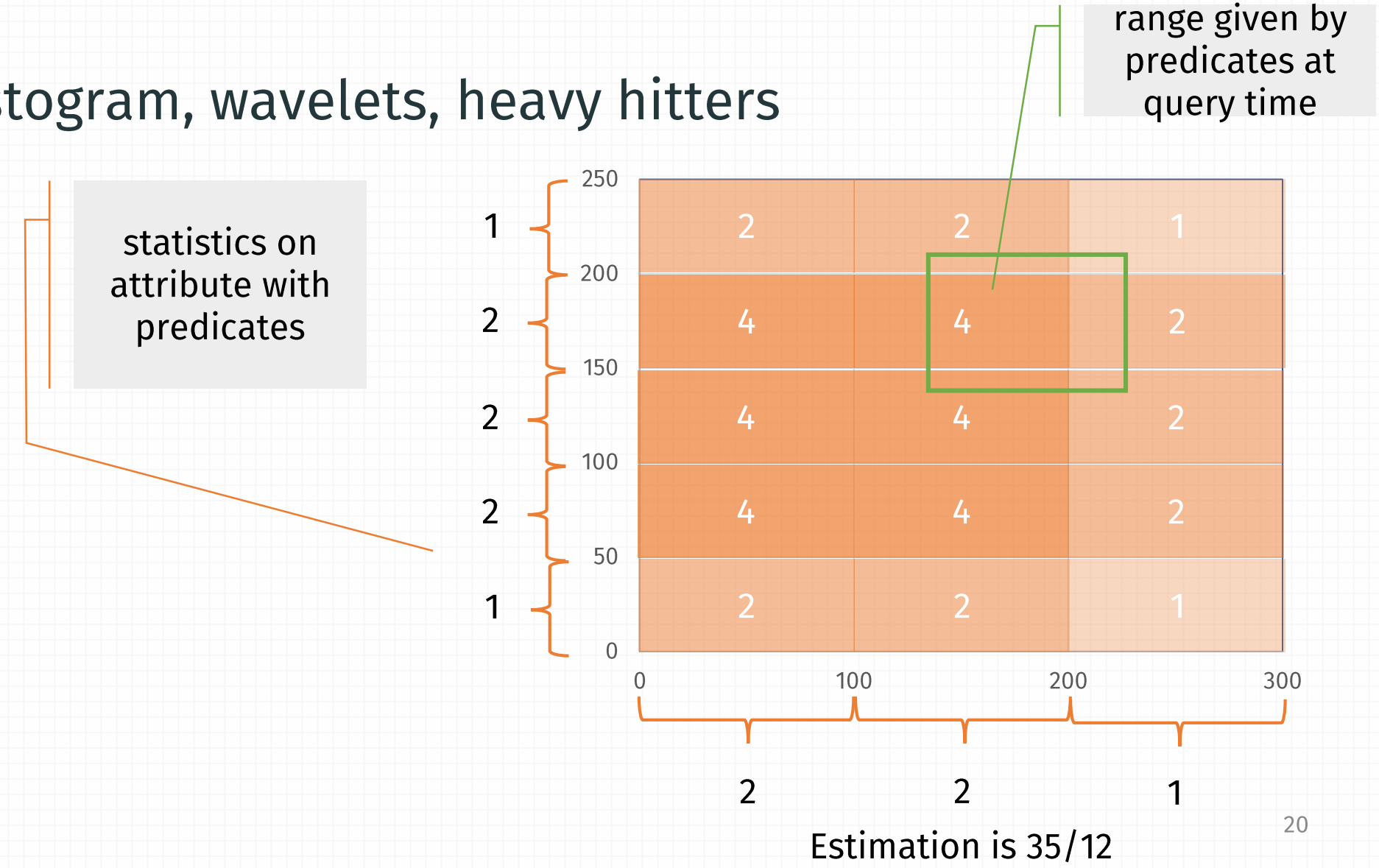
Synopsis

- statistics, histogram, wavelets, heavy hitters



Synopsis

- statistics, histogram, wavelets, heavy hitters



Synopsis

- statistics, histogram, wavelets, heavy hitters
  - only support range/equality predicates
  - sparsity(esp. in high dimension and/or large domain
    - 4-dimension, domain size = 64 [Chakrabarti et al. 2001]

Main approaches

- Sketching
- Synopsis
- Random sampling

Sampling

- “All-purpose” approach
  - For any predicate c_A on A, $sample\left(\sigma_{c_A}(A)\right) = \sigma_{c_A}\left(sample(A)\right)$
- Collect samples offline and give estimation at query time.
- “insensitive” to # columns, domain size, form of predicates, sparsity, etc.

independent Bernoulli Sampling [Vengerov'15]

- Each tuple is sampled independently with probability q.

A

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22

q = 1/2

$\bowtie$

B

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418

$A \bowtie B$

Ord key	Cust key	Name	...
1	2	Elliott	...
2	4	Parker	...
3	5	Wilford	...
4	3	Helga	...
5	3	Helga	...
6	1	Lizabeth	...
7	2	Elliott	...

independent Bernoulli Sampling

- Each tuple is sampled independently with probability q.
- Estimate join size by

$$\hat{J}_{Ber} = \frac{|S_A \bowtie S_B|}{q^2}.$$

Each tuple is sampled with probability ¼ (not independent)

S_A

Cust key	Name	...	Age
1	Lizabeth	...	41
3	Helga	...	20

q = 1/2

$\bowtie$

S_B

Ord key	Cust key	Total Price
1	2	322
3	5	420
6	1	604
7	2	418

$S_A \bowtie S_B$

Ord key	Cust key	Name	...
6	1	Lizabeth	...

Estimation is 4

independent Bernoulli Sampling

- Each tuple is sampled independently with probability q.
- Estimate join size by

$$\hat{J}_{Ber} = \frac{|S_A \bowtie S_B|}{q^2}.$$

- The variance is

$$\mathrm{Var}[\hat{J}_{Ber}] = \sum_v a_v b_v \left[\left(\frac{1}{q^2} - 1 \right) + (a_v - 1) \left(\frac{1}{q} - 1 \right) + (b_v - 1) \left(\frac{1}{q} - 1 \right) \right].$$

- Problem: Ignore the join relationship between the two tables.

Correlated Sampling [Vengerov et.al 15]

- Let $h: [u] \rightarrow [0,1]$ be a random hash function.
- Every tuple with join attribute value v is taken into the sample if $h(v) < p$.

A

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22

B

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418

$A \bowtie B$

Ord key	Cust key	Name	...
1	2	Elliott	...
2	4	Parker	...
3	5	Wilford	...
4	3	Helga	...
5	3	Helga	...
6	1	Lizabeth	...
7	2	Elliott	...

$p = \frac{1}{2}$ $h(1), h(3) < \frac{1}{2}$

Correlated Sampling

- Let $h: [u] \rightarrow [0,1]$ be a random hash function.
- Every tuple with join attribute value v is taken into the sample if $h(v) < p$.
- The estimator is $\hat{J}_{Cor} = \frac{|S_A \bowtie S_B|}{p}$.

Each set of tuples is sampled with probability ½.

S_A

Cust key	Name	…	Age
1	Lizabeth	…	41
3	Helga	…	20

S_B

Ord key	Cust key	Total Price
4	3	82
5	3	120
6	1	604

$S_A \bowtie S_B$

Estimation is 6.

Ord key	Cust key	Name	…
4	3	Helga	…
5	3	Helga	…
6	1	Lizabeth	…

$p = \frac{1}{2}$ $h(1), h(3) < \frac{1}{2}$

Correlated Sampling

- Let $h: [u] \to [0,1]$ be a random hash function.
- Every tuple with join attribute value v is taken into the sample if $h(v) < p$.
- The estimator is $$\hat{J}_{Cor} = \frac{|S_A \bowtie S_B|}{p}.$$

- The variance is $$\mathrm{Var}[\hat{J}_{Cor}] = \left(\frac{1}{p} - 1\right) \sum_v a_v^2 b_v^2.$$

Comparison

- Extreme example:

$$\mathrm{Var}[\hat{J}_{Ber}] \approx \|a\|_1/q^2$$

$$\mathrm{Var}[\hat{J}_{Cor}] \approx \|a\|_1/p$$

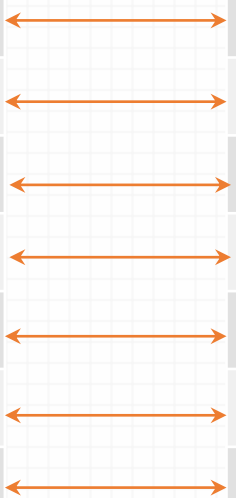
- Bernoulli Sampling has trouble matching the tuples.

A

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22
6	Freddie	...	34
7	Vanessa	...	51

B

Ord key	Cust key	Total Price
4	1	322
2	2	553
1	3	420
7	4	82
5	5	120
3	6	604
6	7	418



one-to-one

Comparison

- Extreme example:

$$\mathrm{Var}[\hat{J}_{Ber}] \approx \|a\|_1\|b\|_1/q^2$$

$$\mathrm{Var}[\hat{J}_{Cor}] \approx \|a\|_1^2\|b\|_1^2/p$$

- Correlated Sampling takes all or takes none.

A

Follower	Followee
1	3
2	3
4	3
5	3
6	3
7	3
8	3

B

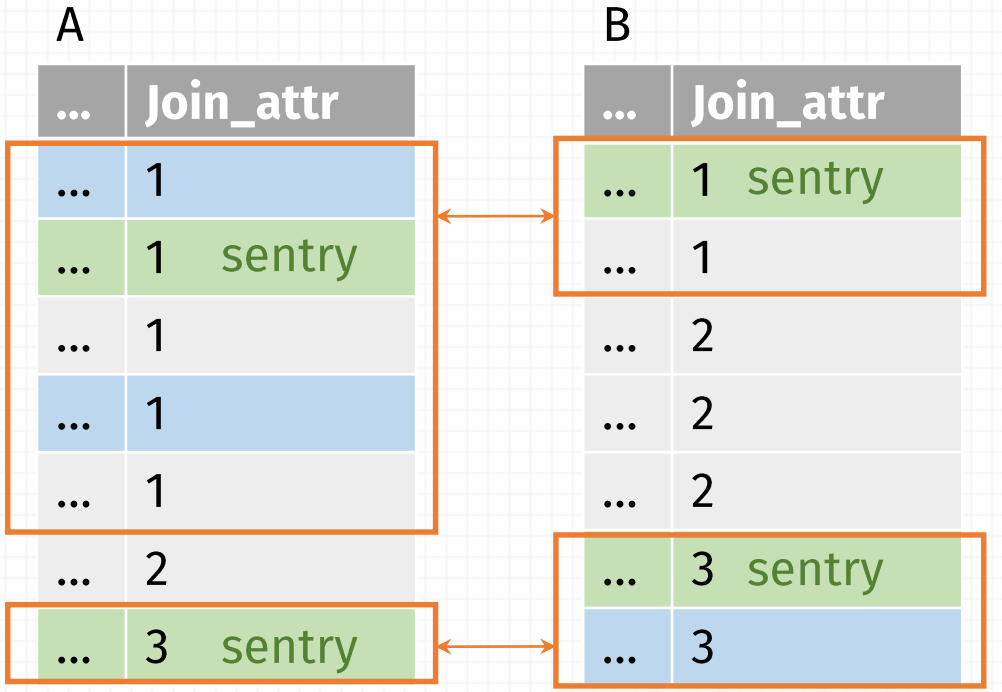
Follower	Followee
1	3
2	3
4	3
5	3
6	3
7	3
8	3

Cartesian Product

Two-Level Sampling

Two-Level Sampling

- Let $h : [u] \rightarrow [0, 1]$ be a random hash function.
- Level 1:
  - For each join value v with $h(v) < p$,
  - one tuple in $A(v)$ is chosen uniformly at random as the '*sentry*'
- Level 2:
  - Each of the rest tuples in $A(v)$ is sampled independently with probability q.
  - Denoted as set $S_A(v)$.
- The same is done for table B.



$$p = \frac{1}{2} \quad q = \frac{1}{2}$$

$$h(1), h(3) < \frac{1}{2}$$

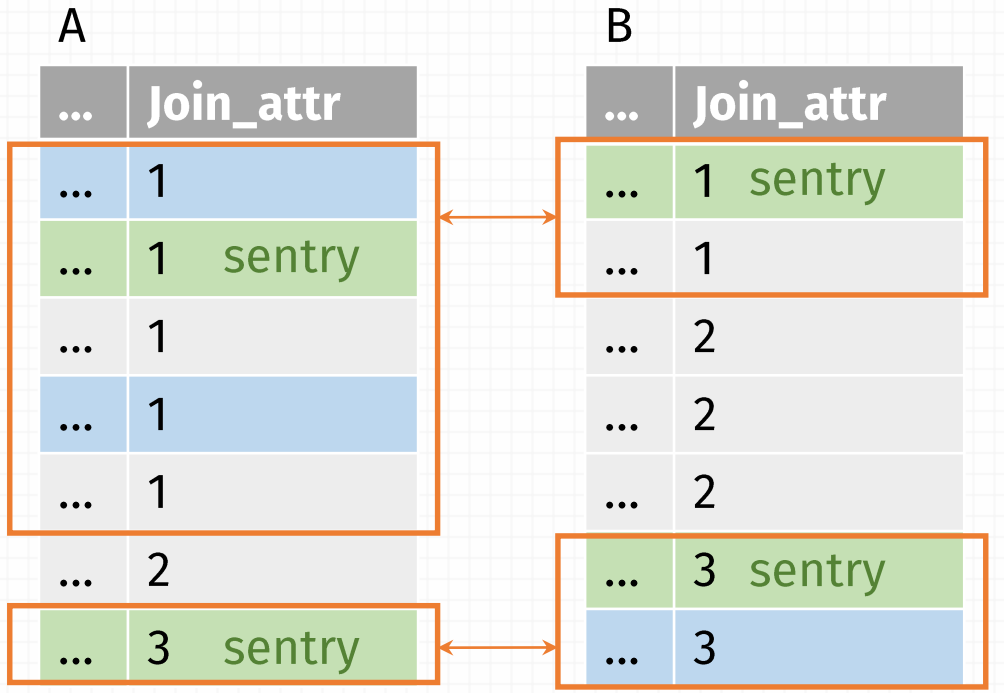
Two-Level Sampling

- An unbiased estimator for $J_v = a_v b_v$ is

$$\hat{J}_v = \begin{cases} \frac{1}{p}\left(\frac{|S_A(v)|}{q}+1\right)\left(\frac{|S_B(v)|}{q}+1\right), & \text{if } v \text{ is sampled for } A \text{ and } B; \\ 0, & \text{otherwise.} \end{cases}$$

- Estimate J by $\hat{J}_{2lvl} = \Sigma_v \hat{J}_v$.

$S_A(v)$ and $S_B(v)$ are not independent. They are independent conditioned on $h(v) < p$.



$p = \frac{1}{2}$ $q = \frac{1}{2}$

$h(1), h(3) < \frac{1}{2}$

Two-Level Sampling

- An unbiased estimator for $J_v = a_v b_v$ is

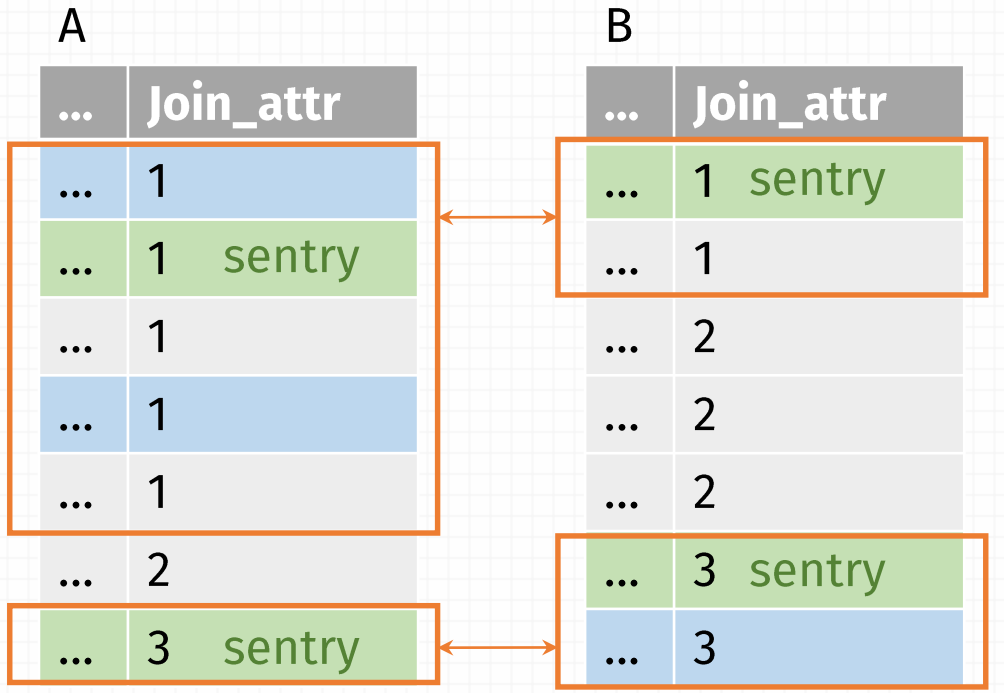
- An unbiased estimator for is

- Estimate J by .

- Estimate J by Conditioned on v is sampled: $E(|S_A(v)|) = a_v - 1$, $E(|S_B(v)|) = b_v - 1$.

Conditioned on v is sampled:

$$E(|S_A(v)|) = a_v - 1$$
$$E(|S_B(v)|) = b_v - 1$$



$$p = \frac{1}{2} \quad q = \frac{1}{2}$$

$$h(1), h(3) < \frac{1}{2}$$

Two-Level Sampling

- The variance is

$$\mathrm{Var}[\hat{J}_{2lvl}] = \Sigma_{v:a_v b_v \neq 0} \mathrm{Var}[\hat{J}_v] = \Sigma_{v:a_v b_v \neq 0} \left(\frac{1}{p}\sigma_1^2(v) + \sigma_2^2(v) \right),$$

where

$$\sigma_1^2(v) = \left(\frac{1}{q^2} - 1\right)(a_v - 1)(b_v - 1) + \left(\frac{1}{q} - 1\right)(b_v - 1)(a_v^2 - a_v + 1) + \left(\frac{1}{q} - 1\right)(a_v - 1)(b_v^2 - b_v + 1),$$

$$\sigma_2^2(v) = \left(\tfrac{1}{p} - 1\right) a_v^2 b_v^2.$$

Two-Level Sampling

Two-level sampling(A)

```
1  for every tuple t in A do
2      if h(t.v) < p then
3          if v has never been sampled then
4              s_A(v) ← t
5              c_v ← 1
6          else
7              c_v ← c_v + 1
8              set s_A(v) ← t with probability 1/c_v
9          Sample t with probability q into S_A(t.v)
10 Exclude s_A(v) from S_A(v) for each v
```

Use Reservoir Sampling to sample the sentry

The algorithm can be easily performed in one pass over the data

Two questions

- How to set the parameters p and q?

$$\begin{aligned} \text{minimize} \quad & \mathrm{Var}[\hat{J}_{2lvl}] \\ \text{s.t.} \quad & \text{sample size} = n \\ & p, q \in [0, 1]. \end{aligned}$$

  - Assume given basic statistics: $\|a\|_0, \|a\|_1, \|b\|_0, \|b\|_1$ ...

- Compare the variance with the existing algorithms while the sampling size n is fixed.

Comparison

- PK-FK joins: $b_v = 1$
- Many-many joins

PK-FK Joins

- Optimal p and q can be found.

$$q = \begin{cases} \frac{n-\|a\|_0-\|b\|_1}{\|a\|_1-\|a\|_0}, & \text{if } \sqrt{\frac{\|a\|_0+\|b\|_1}{\|a\|_2^2-\|a\|_1+\|a\|_0}} < \frac{n-\|a\|_0-\|b\|_1}{\|a\|_1-\|a\|_0}; \\ 1, & \text{if } \sqrt{\frac{\|a\|_0+\|b\|_1}{\|a\|_2^2-\|a\|_1+\|a\|_0}} > 1; \\ \sqrt{\frac{\|a\|_0+\|b\|_1}{\|a\|_2^2-\|a\|_1+\|a\|_0}}, & \text{otherwise,} \end{cases}$$

and

$$p = \frac{n}{\|b\|_1 + \|a\|_0 + q(\|a\|_1 - \|a\|_0)}.$$

PK-FK Joins

Theorem

$$\mathrm{Var}[\hat{J_{2lvl}}] \leq \mathrm{Var}[\hat{J_{Cor}}] < \mathrm{Var}[\hat{J_{Ber}}]$$

PK-FK Joins

- Two-Level Sampling is always better than Correlated Sampling.

The larger the FK table, the better

$$\frac{\mathrm{Var}[\hat{J_{2lvl}}]}{\mathrm{Var}[\hat{J_{Cor}}]} \approx \frac{\|b\|_1}{\|a\|_1} + \sqrt{\frac{\|b\|_1}{\|a\|_2^2}}.$$

The higher skew, the better

PK-FK Joins

- Two-Level Sampling is always better than Correlated Sampling.

$$\frac{\mathrm{Var}[\hat{J_{2lvl}}]}{\mathrm{Var}[\hat{J_{Cor}}]} \approx \frac{\|b\|_1}{\|a\|_1} + \sqrt{\frac{\|b\|_1}{\|a\|_2^2}}.$$

The larger the FK table, the better

The higher skew, the better

- Note that we usually have

$$\|a\|_2^2 \gg \|a\|_1 \gg \|b\|_1$$

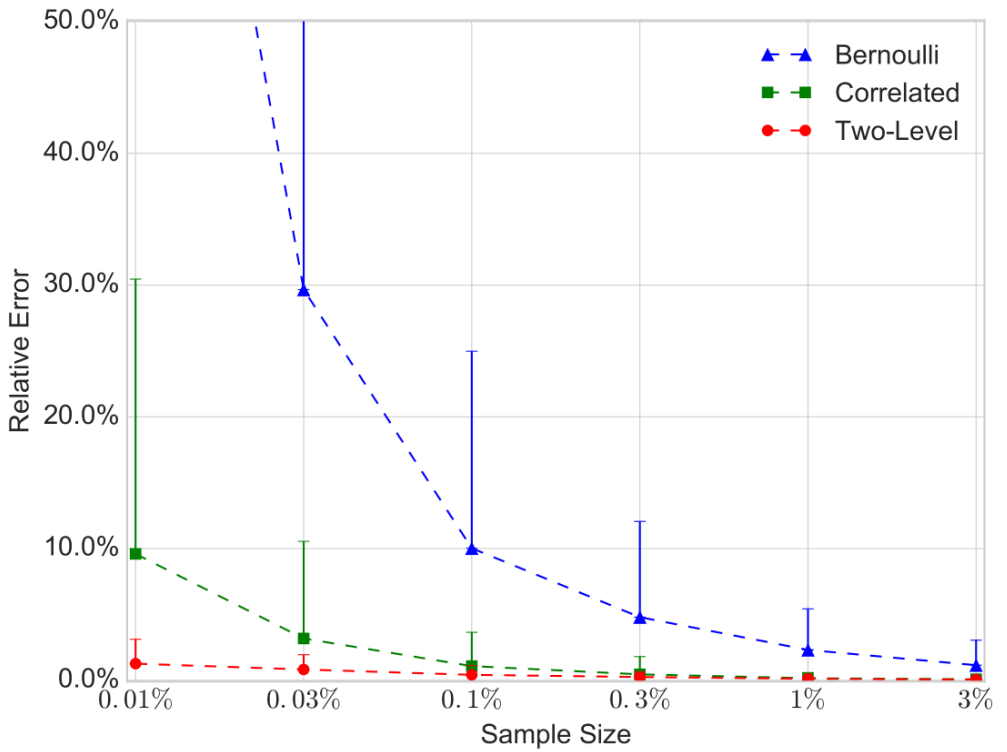
$$\|a\|_1 + \|b\|_1 \gg n$$

Experiments

- PK-FK join on TPC-H data
  - 500 times
  - relative error

$$\frac{|J - \hat{J}|}{J}$$

  - median error and the 90%-quantile error



PK-FK join on TPC-H data.

$$\|a\|_1 = 5.9 \times 10^7, \|a\|_0 = \|b\|_1 = 10^5, \|a\|_2^2 = 3.6 \times 10^{10}$$

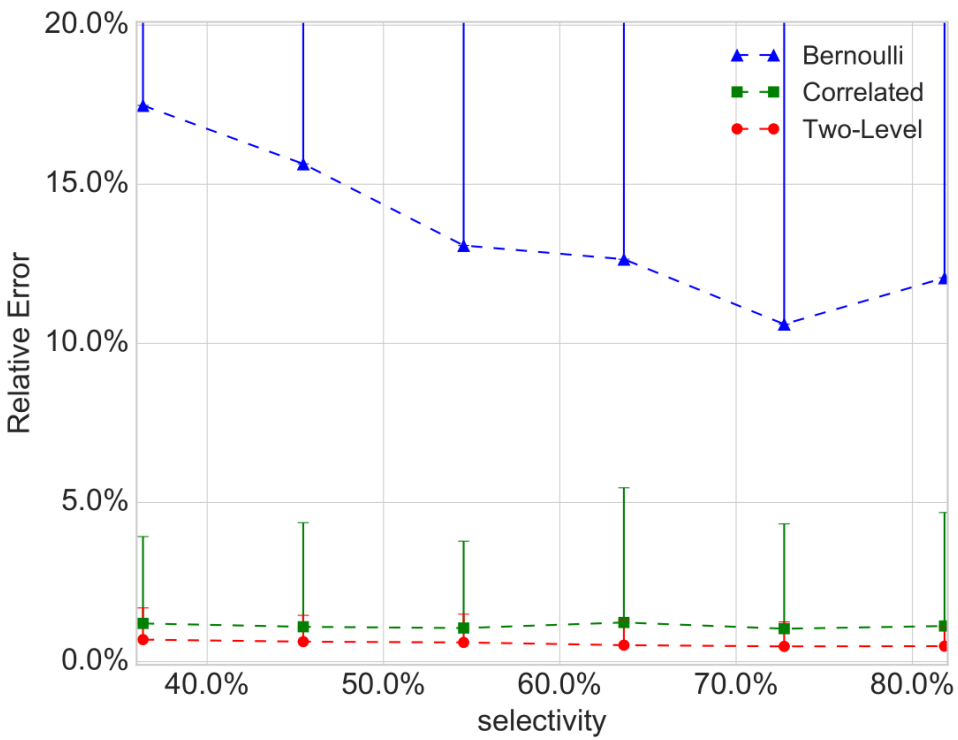
Experiments

- PK-FK join on TPC-H data
  - 500 times
  - relative error

$$\frac{|J - \hat{J}|}{J}$$

  - median error and the 90%-quantile error

  - 1 predicate
  - sample size is fixed at 0.1%



PK-FK join on TPC-H data with 1 predicate

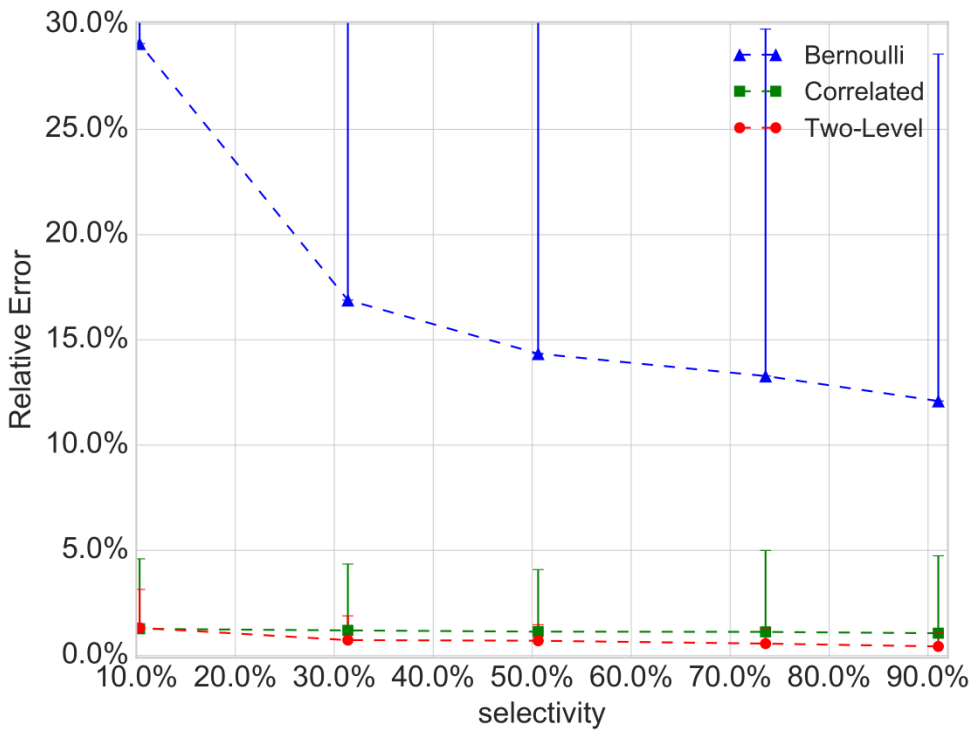
$$\|a\|_1 = 5.9 \times 10^7, \|a\|_0 = \|b\|_1 = 10^5, \|a\|_2^2 = 3.6 \times 10^{10}$$

Experiments

- PK-FK join on TPC-H data
  - 500 times
  - relative error

$$\frac{|J - \hat{J}|}{J}$$

  - median error and the 90%-quantile error
  - 2 predicates
  - sample size is fixed at 0.1%

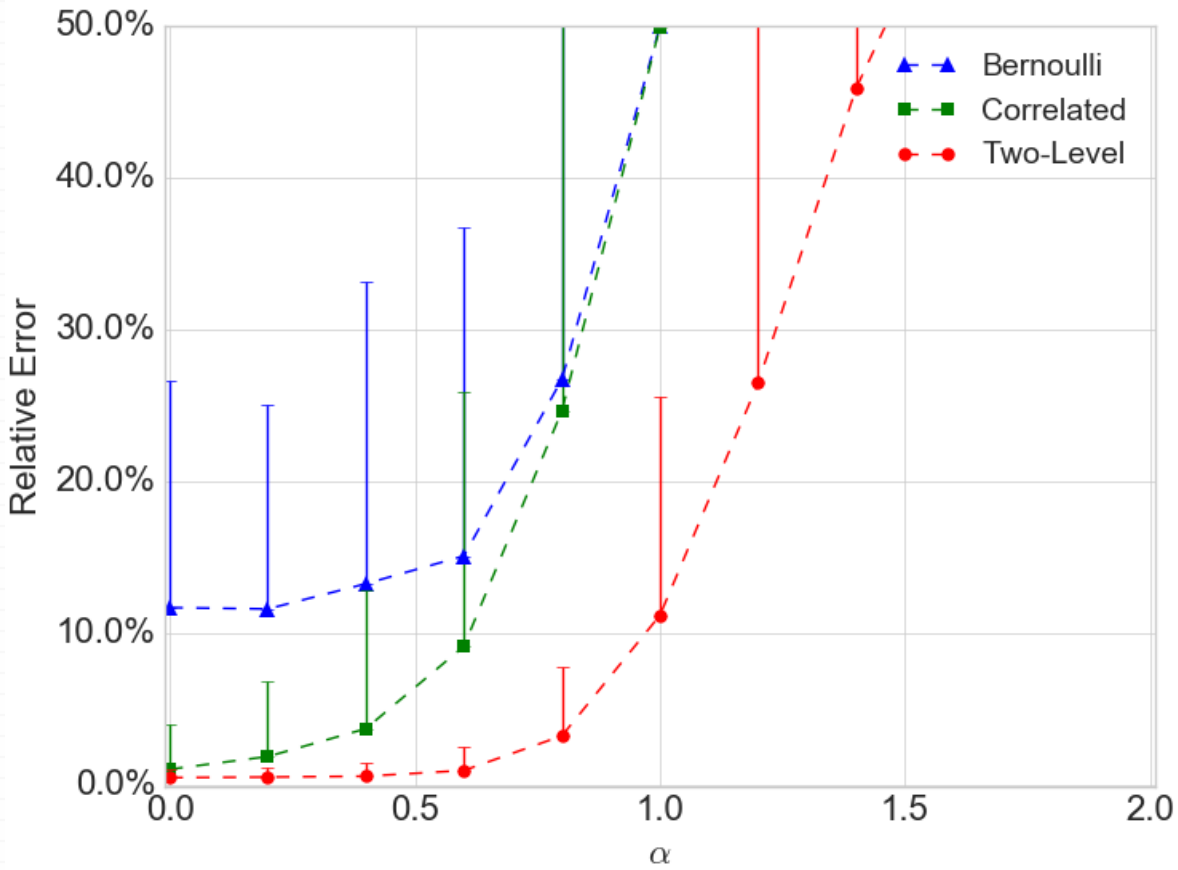


PK-FK join on TPC-H data with 2 predicates

$$\|a\|_1 = 5.9 \times 10^7, \|a\|_0 = \|b\|_1 = 10^5, \|a\|_2^2 = 3.6 \times 10^{10}$$

Experiments

- PK-FK join on skewed data
  - Zipf distribution
  - Skewness is controlled by the Zipf parameter α.
  - sample size is fixed at 0.1%



PK-FK join on skewed data

Many-many Joins

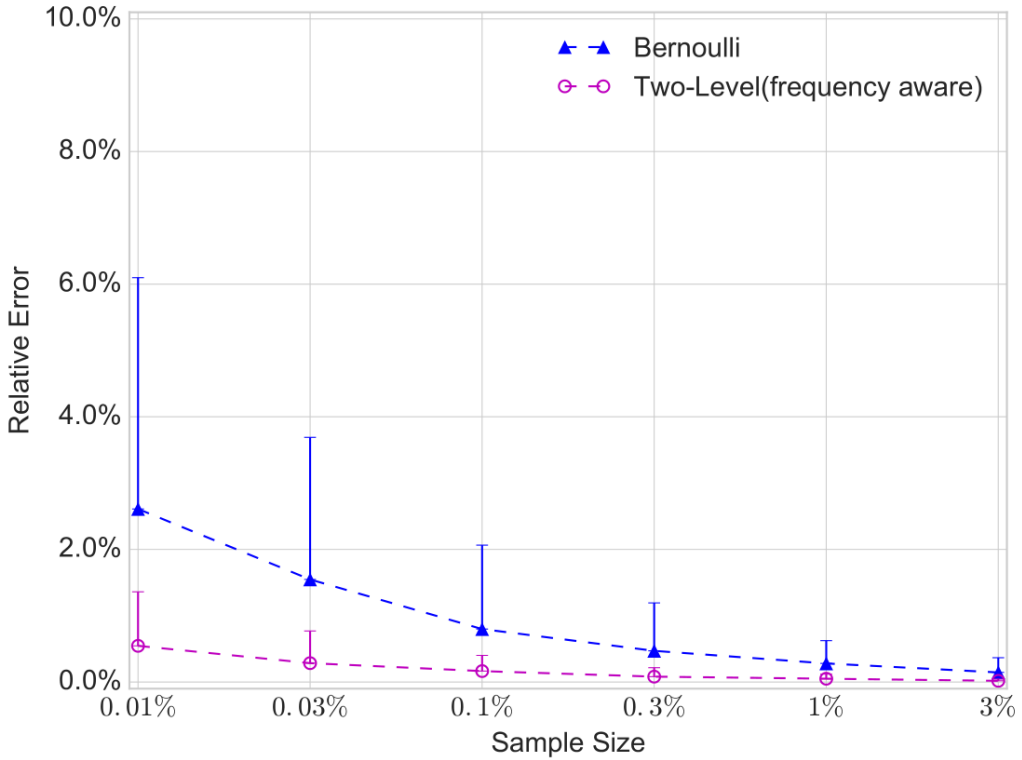
- Optimal p and q can be found by solving a degree-4 equation.

$$\frac{\mathrm{Var}[\hat{J}_{2lvl}]}{\mathrm{Var}[\hat{J}_{Cor}]} \approx \frac{\sqrt{\|a \circ b\|_1} \cdot \|a \circ b\|_2 + \|a \circ b\|_2^{\frac{3}{2}}}{\|a \circ b\|_2^2} \leq 1 \text{ asymptotically,}$$

where $x \circ y$ is defined as the vector $(x_1 y_1, x_2 y_2, \ldots)$.

Experiments

- Many-many join on twitter data



Many-many join on twitter data

1.4×10^9 tuples involving 4.2×10^7 users

Frequency-aware Sampling

- Use frequency information to improve the accuracy of sampling.

End-biased Sampling [Estan et.cl 2006]

- “Frequency-aware” version of Correlated Sampling
- Favor join value with higher frequency

End-biased Sampling

- Let $h: [u] \rightarrow [0,1]$ be a random hash function.
- Every tuple with join attribute value v is taken into the sample if $h(v) < p_v$, where $p_v \propto a_v$.

$p_v = 1/2$

A

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22

B

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	2	82
5	2	120
6	1	604
7	2	418

$p_2 = 1$

$p_v = 1/4$ for $v \neq 2$

frequency-aware Two-Level Sampling

- Use p_v as the first level sampling probability for each v.
- Optimal q can be found by numerical methods. Given q:

$$p_v = C \cdot \sqrt{\frac{\sigma_1^2(v) + a_v^2 b_v^2}{2 + q(a_v + b_v - 2)}}$$

- Setting q to 1, p_v should be proportional to $\frac{a_v b_v}{\sqrt{a_v + b_v}}$.

Comparison

- PK-FK Joins:
  - $p_v \propto \sqrt{a_v}$
- Many-many joins:
  - Example:
  - $p_v \propto a_v^{1.5}$ when $a_v = b_v$.

Comparison

- PK-FK Joins:
  - $p_v \propto \sqrt{a_v}$
- Many-many joins:
  - Example:
  - $p_v \propto a_v^{1.5}$ when $a_v = b_v$.

Intuition:
Tuples with value v contribute a_v to samples while they contribute a_v^2 to the variance.

Comparison

- PK-FK Joins:
  - $p_v \propto \sqrt{a_v}$

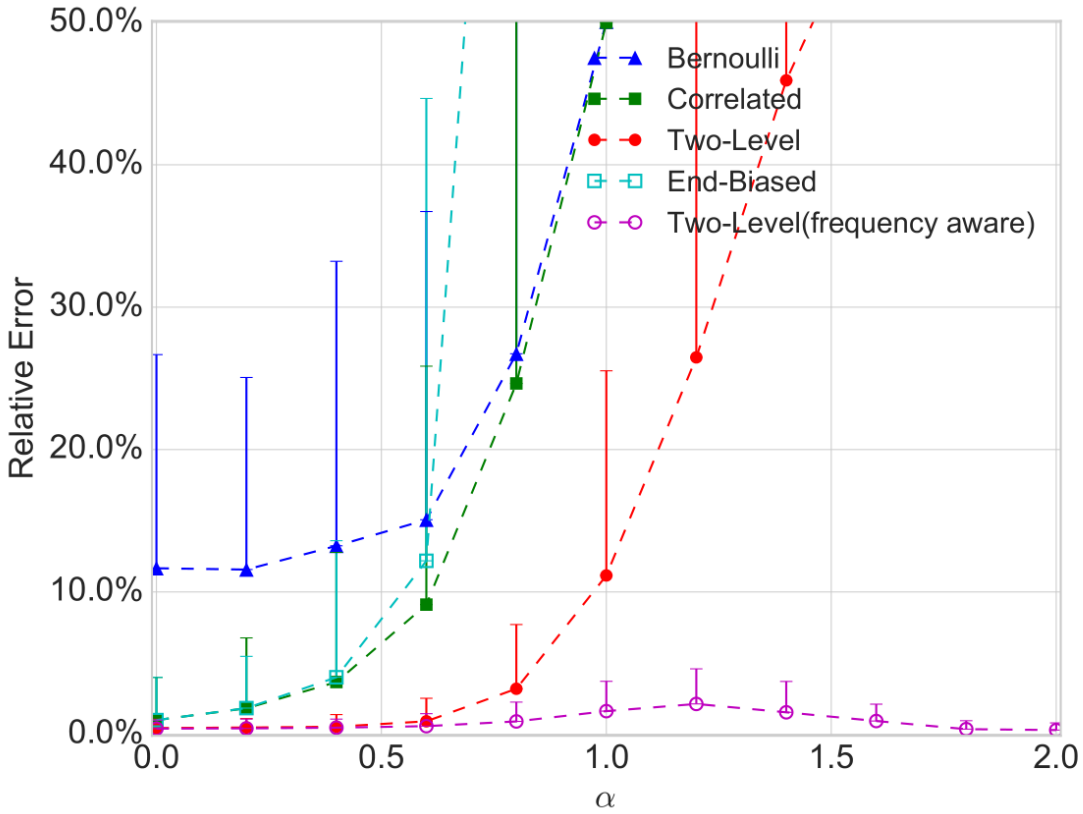
Intuition:
Tuples with value v contribute a_v to samples while they contribute a_v^2 to the variance.

- Many-many joins:
  - Example:
  - $p_v \propto a_v^{1.5}$ when $a_v = b_v$.

Intuition:
Tuples with value v contribute a_v to samples while they contribute a_v^4 to the variance.

Experiments

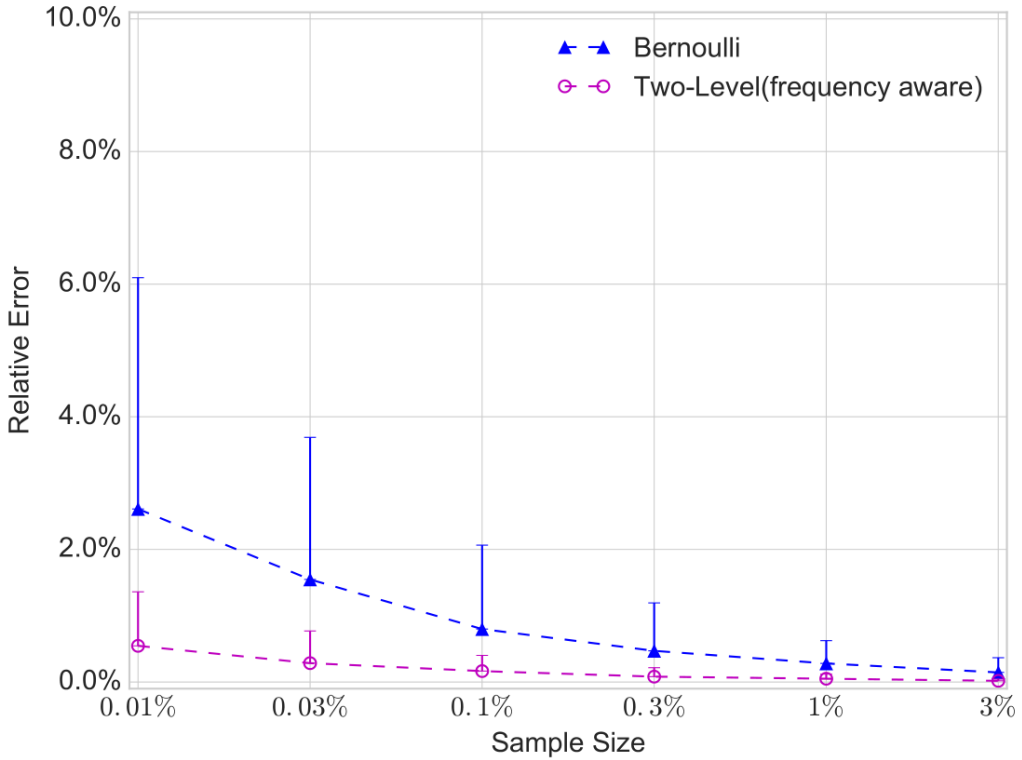
- PK-FK join on skewed data
  - Zipf distribution
  - Skewness is controlled by the Zipf parameter α.
  - sample size is fixed at 0.1%



PK-FK join on skewed data

Experiments

- Many-many join on twitter data



Many-many join on twitter data

1.4×10^9 tuples involving 4.2×10^7 users

Summary

- Build upon and extend three sampling algorithms
  - Bernoŭlli sampling
  - Correlated sampling
  - End-biased sampling
- Achieve smaller variance for both PK-FK joins and many-to-many joins

- Extension:
  - Give confidence interval, extend to SUM and AVG queries
  - Extend to chain joins and star joins

Questions?

References

[Rusu et.al 2008] F. Rusu and A. Dobra. **Sketches for size of join estimation**. *ACM Transactions on Database Systems, 33:1-46, 2008*.

[Vengerov et.al 2015] D. Vengerov, A. C. Menck, M. Zait, and S. P. Chakkappen. **Join size estimation subject to filter conditions**. *Proceedings of the VLDB Endowment, 8(12):1530-1541, 2015*.

[Estan et.al 2006] C. Estan and J. F. Naughton. **End-biased samples for join cardinality estimation**. *In Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on, pages 20{20. IEEE, 2006*.

[Chakrabarti et.al 2001] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim. **Approximate query processing using wavelets**. *The VLDB Journal, 10(2-3):199{223, 2001*.

[Dobra et.cl 2002] A. Dobra, M. Garofalakis, J. Gehrke, and R. Rastogi**. Processing complex aggregate queries over data streams**. *In Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pages 61-72. ACM, 2002*.

ppt template from mtheme (https://github.com/matze/mtheme)

PK-FK Joins

- Two-Level Sampling is always better than Correlated Sampling.

The larger the FK table, the better

$$\frac{\mathrm{Var}[\hat{J_{2lvl}}]}{\mathrm{Var}[\hat{J_{Cor}}]} \approx \frac{\|b\|_1}{\|a\|_1} + \sqrt{\frac{\|b\|_1}{\|a\|_2^2}}.$$

- When sample size $n \geq \|a\|_0 + \|b\|_0$

$$\frac{\mathrm{Var}[\hat{J}_{2lvl}]}{\mathrm{Var}[\hat{J}_{Cor}]} < \frac{n}{\|a\|_1 + \|b\|_1}.$$

The higher skew, the better

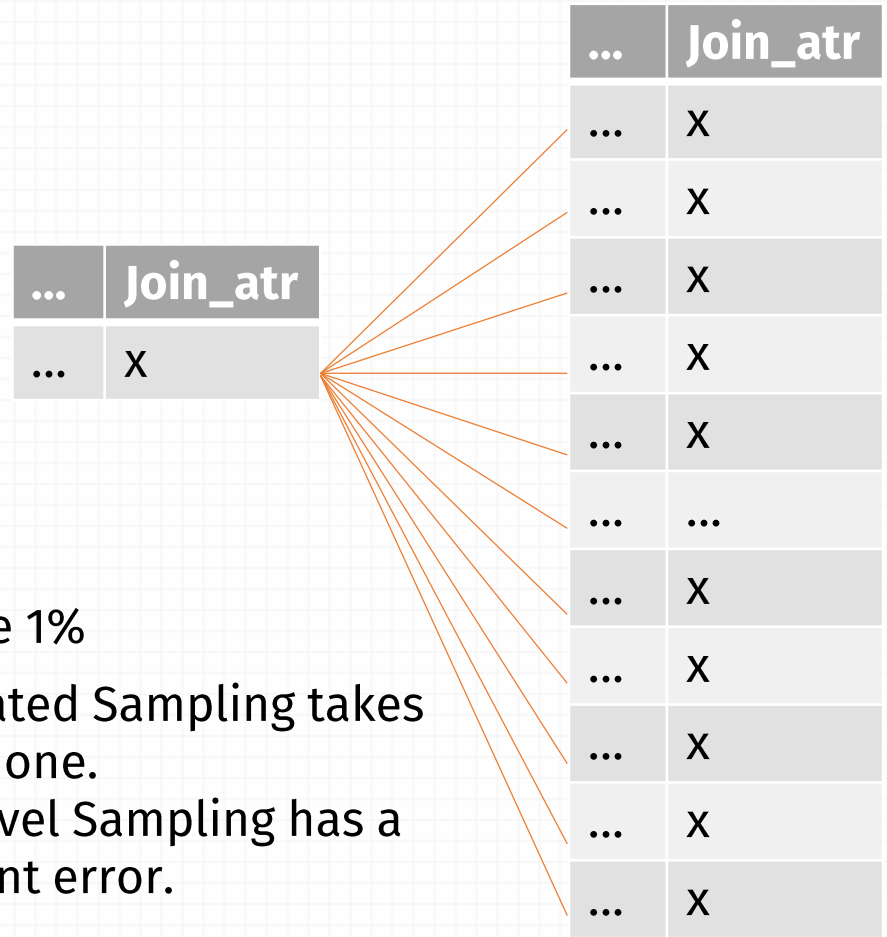
- Note that we usually have $\|a\|_2^2 \gg \|a\|_1 \gg \|b\|_1$ and $\|a\|_1 + \|b\|_1 \gg n$.

PK-FK Joins

$$\frac{\mathrm{Var}[\hat{J_{2lvl}}]}{\mathrm{Var}[\hat{J_{Cor}}]} \approx \frac{\|b\|_1}{\|a\|_1} + \sqrt{\frac{\|b\|_1}{\|a\|_2^2}}.$$

When sample size $n \geq \|a\|_0 + \|b\|_0$

$$\frac{\mathrm{Var}[\hat{J}_{2lvl}]}{\mathrm{Var}[\hat{J}_{Cor}]} < \frac{n}{\|a\|_1 + \|b\|_1}.$$



Sample 1%

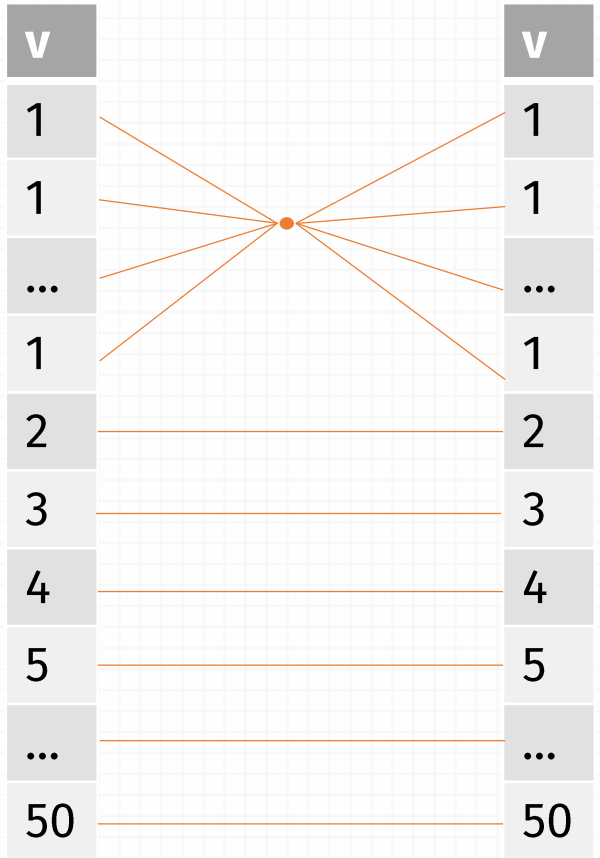
Correlated Sampling takes all or none.
Two-level Sampling has a constant error.

Many-many Joins

- There is no definitive comparison between two-level sampling and independent Bernoulli sampling.

Many-many Joins

- There is no definitive comparison between two-level sampling and independent Bernoulli sampling.
- Two-Level sampling suffers by
  - giving same weight for all v on first level

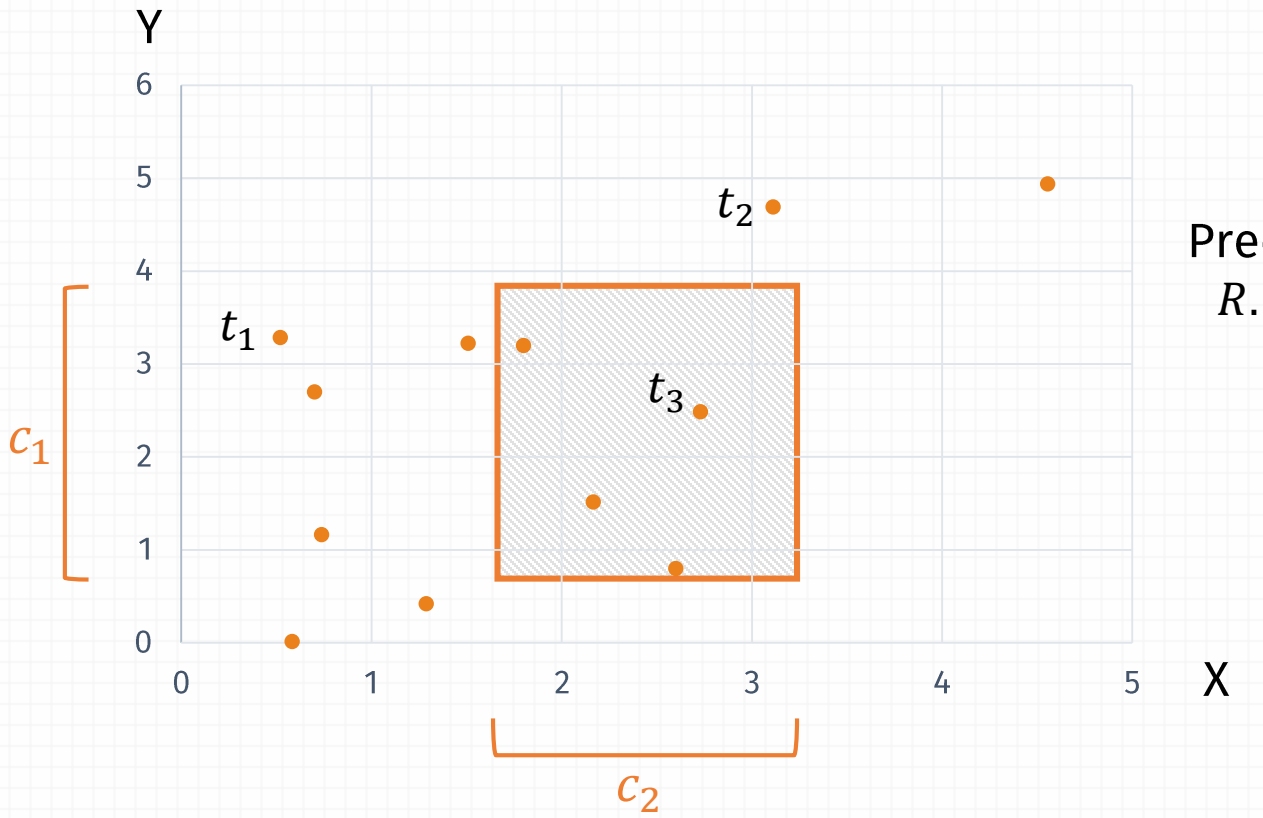


Synopsis

- statistics, histogram, wavelets, heavy hitters
  - predicates: based on hierarchical decomposition of the multi-dimensional space
- Example:

R

	x	y
t_1	…	…
t_2	…	…
t_3	…	…
	…	…
	…	…



Predicates:

$R.x \in c_1$ and $R.y \in c_2$

- **Problem: Estimate join size**
  - **Selection predicates given at query time**

$$\text{Customer} \bowtie \text{Order}$$

$$\sigma_{\text{Age}<=35}(Customer) \bowtie \sigma_{\text{TotalPrice}>200}(Order)$$

- **Our solution: Sampling**

$$\text{sample}(\sigma_c(R)) = \sigma_c(\text{sample}(R))$$

- **Two-Level Sampling**
- **One Pass, Unbiased, Smaller Error**
- **Beats previous sampling methods**

Customer

Cust key	Name	...	Age
1	Lizabeth	...	41
2	Elliott	...	65
3	Helga	...	20
4	Parker	...	47
5	Wilford	...	22

Order

Ord key	Cust key	Total Price
1	2	322
2	4	553
3	5	420
4	3	82
5	3	120
6	1	604
7	2	418