Pessimistic Query Optimization: Tighter Upper Bounds for Intermediate Join Cardinalities

Walter Cai   Magdalena Balazinska   Dan Suciu

University of Washington

[walter,magda,suciu]@cs.washington.edu

July 19th, 2019

Systematic Underestimation

Query optimizers assume:

- Uniformity
- Independence

1. Background: Cardinality Bounds
2. Tightened Cardinality Bounds
3. Evaluation

Example Query (SQL)

```
SELECT
    *
FROM
    pseudonym,
    cast,
    movie_companies,
    company_name
WHERE
    pseudonym.person_id = cast.person_id AND
    cast.movie_id_id = movie_companies.movie_id AND
    movie_companies.company_id = company_name.id;
```

Example Query (Join Graph & Datalog)

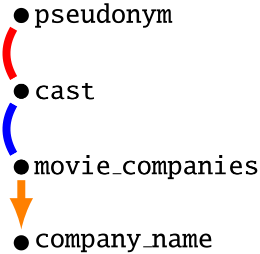


Figure: Join Graph.

$$Q(x, y, z, w) :\!- pseudo(x, y), \; cast(y, z), \; mc(z, w), \; cn(w)$$

$y \mapsto$ person

$z \mapsto$ movie

$w \mapsto$ company

Review: Entropy

Take random variable X:

$$h(X) = -\sum_{a} \mathbb{P}(X = a) \cdot \log(\mathbb{P}(X = a))$$

Multiple variables:

$$h(X, Y) = -\sum_{a,b} \mathbb{P}(X = a, Y = b) \cdot \log(\mathbb{P}(X = a, Y = b))$$

Conditional entropy:

$$h(X|Y) = -\sum_{a,b} \mathbb{P}(X = a, Y = b) \cdot \log\left(\frac{\mathbb{P}(X = a, Y = b)}{\mathbb{P}(Y = b)}\right)$$

Review: Entropy

$$X \sim P(x_1, \ldots, x_n)$$

- Fact: $h(X) \leq \log(n)$
- $h(X) = \log(n)$ iff P is uniform

Connection to Entropy

$$Q(x, y, z, w) \text{ :- } pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$

- Create random variable for each attribute.

$$x \to X, \quad y \to Y, \quad z \to Z, \quad w \to W$$

- Let (X, Y, Z, W) be uniformly distributed over true output of Q.

$$h(X, Y, Z, W) = \log \left| Q(x, y, z, w) \right|$$
$$\exp\left(h(X, Y, Z, W)\right) = \left| Q(x, y, z, w) \right|$$

Entropic Bound

$$\left|Q(x, y, z, w)\right| = \exp\left(h(X, Y, Z, W)\right)$$

- Suffices to bound $h(X, Y, Z, W)$.

Entropic Bound

$$h(X, Y, Z, W) \leq h(X|Y) + h(Y, Z) + h(W|Z)$$

Entropic Bound

$$\left|Q(x,y,z,w)\right| = \exp(h(X,Y,Z,W)) \le \exp(h(X|Y) + h(Y,Z) + h(W|Z))$$

$$h(Y,Z) \le \log(\mathtt{count}(\mathtt{cast}))$$
$$h(X|Y) \le \log(\mathtt{max\ degree}(\mathtt{pseudonym}))$$
$$h(W|Z) \le \log(\mathtt{max\ degree}(\mathtt{movie\_companies}))$$

Entropic Bound

$$\left|Q(x, y, z, w)\right| = \exp(h(X, Y, Z, W)) \leq \exp(h(X|Y) + h(Y, Z) + h(W|Z))$$

$$h(Y, Z) \leq \log c_{\texttt{cast}}$$
$$h(X|Y) \leq \log d^{y}_{\texttt{pseudo}}$$
$$h(W|Z) \leq \log d^{z}_{\texttt{mc}}$$

Cardinality Bound

$$
\begin{aligned}
\left|Q(x,y,z,w)\right| &= \exp(h(X,Y,Z,W)) \\
&\leq \exp(\underbrace{h(X|Y)}_{\leq \log d^{y}_{\mathtt{pseudo}}} + \underbrace{h(Y,Z)}_{\leq \log c_{\mathtt{cast}}} + \underbrace{h(W|Z)}_{\leq \log d^{z}_{\mathtt{mc}}}) \\
&\leq d^{y}_{\mathtt{pseudo}} \cdot c_{\mathtt{cast}} \cdot d^{z}_{\mathtt{mc}}
\end{aligned}
$$

Many Entropic Bounds

$h(X, Y, Z, W) \leq ...$
$h(X, Y) + h(Z\|Y) + h(W\|Z)$
$h(X, Y) + h(Z\|Y) + h(W)$
$h(X, Y) + h(Z, W)$
$h(X, Y) + h(Z\|W) + h(W)$
$h(X\|Y) + h(Y, Z) + h(W\|Z)$
$h(X\|Y) + h(Y\|Z) + h(Z, W)$
$h(X\|Y) + h(Y\|Z) + h(Z\|W) + h(Z)$

Entropic Bounds

$$Q(x, y, z, w) \coloneq pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$

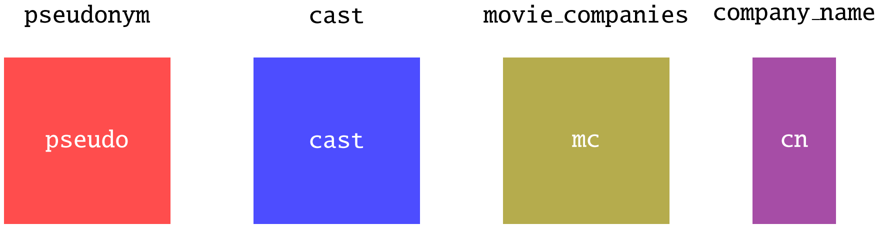
$$\left|Q(x, y, z, w)\right| \leq \min \begin{cases} c_{\texttt{pseudo}} \cdot d^{y}_{\texttt{cast}} \cdot d^{z}_{\texttt{mc}} \\ c_{\texttt{pseudo}} \cdot d^{y}_{\texttt{cast}} \cdot c_{\texttt{cn}} \\ c_{\texttt{pseudo}} \cdot c_{\texttt{mc}} \\ c_{\texttt{pseudo}} \cdot d^{w}_{\texttt{mc}} \cdot c_{\texttt{cn}} \\ d^{y}_{\texttt{pseudo}} \cdot c_{\texttt{cast}} \cdot d^{z}_{\texttt{mc}} \\ d^{y}_{\texttt{pseudo}} \cdot c_{\texttt{cast}} \cdot c_{\texttt{cn}} \\ d^{y}_{\texttt{pseudo}} \cdot d^{z}_{\texttt{cast}} \cdot c_{\texttt{mc}} \\ d^{y}_{\texttt{pseudo}} \cdot d^{z}_{\texttt{cast}} \cdot d^{w}_{\texttt{mc}} \cdot c_{\texttt{cn}} \end{cases}$$

Entropic Bounds

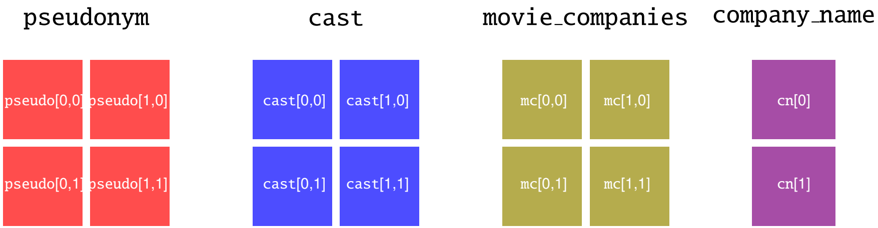
Neat! But is it useful?

- Short answer: ‘No’. (Not yet, anyway)
  - Bounds still too loose (overestimation)
  - Need to tighten
- How to tighten? Partitioning

$$Q(x, y, z, w) :- pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$

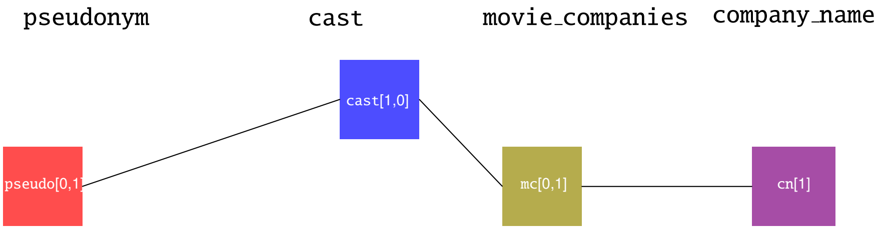


$$Q(x, y, z, w) :- pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$



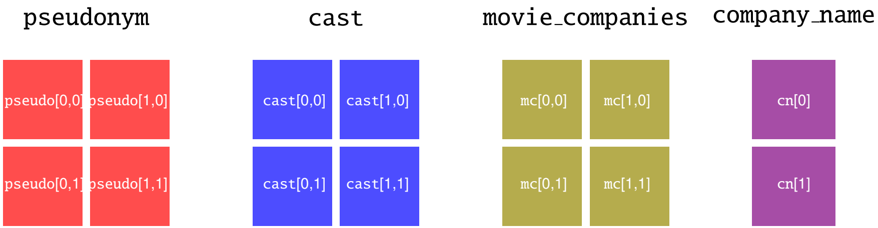
- Value based hashing
- Analagous to hash join

$$Q(x, y, z, w) \text{ :- } pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$



- Value based hashing
- Analagous to hash join

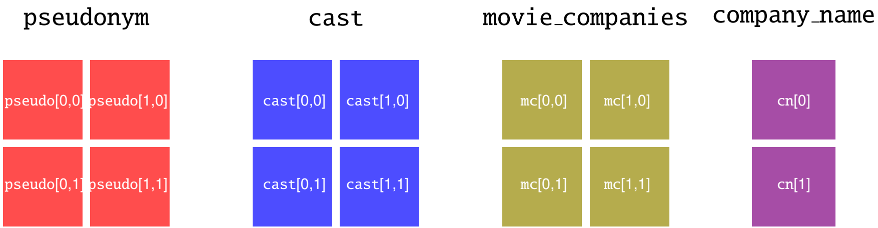
$$Q(x, y, z, w) \mathrel{:\!-} \mathit{pseudo}(x, y), \mathit{cast}(y, z), \mathit{mc}(z, w), \mathit{cn}(w)$$



- $Q(D)$: query evaluated on database D
- $Q(D[J])$: query evaluated on parition $D[J]$

$$Q(D) = \bigcup_J Q(D[J])$$

$$Q(x, y, z, w) :- pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$



- Bound each partition $D[J]$
- Sum will be bound on full database

$$Q(D) = \bigcup_J Q(D[J])$$

$$\left|Q(D)\right| \leq \sum_J bound(Q(D[J]))$$

Partition Bounding

$$\left|Q(D)\right| \leq \sum_{J \in \{0,1\}^4} \min \begin{cases} c_{\mathtt{pseudo}[J]} \cdot d^{y}_{\mathtt{cast}[J]} \cdot d^{z}_{\mathtt{mc}[J]} \\ c_{\mathtt{pseudo}[J]} \cdot d^{y}_{\mathtt{cast}[J]} \cdot c_{\mathtt{cn}[J]} \\ c_{\mathtt{pseudo}[J]} \cdot c_{\mathtt{mc}[J]} \\ c_{\mathtt{pseudo}[J]} \cdot d^{w}_{\mathtt{mc}[J]} \cdot c_{\mathtt{cn}[J]} \\ d^{y}_{\mathtt{pseudo}[J]} \cdot c_{\mathtt{cast}[J]} \cdot d^{z}_{\mathtt{mc}[J]} \\ d^{y}_{\mathtt{pseudo}[J]} \cdot c_{\mathtt{cast}[J]} \cdot c_{\mathtt{cn}[J]} \\ d^{y}_{\mathtt{pseudo}[J]} \cdot d^{z}_{\mathtt{cast}[J]} \cdot c_{\mathtt{mc}[J]} \\ d^{y}_{\mathtt{pseudo}[J]} \cdot d^{z}_{\mathtt{cast}[J]} \cdot d^{w}_{\mathtt{mc}[J]} \cdot c_{\mathtt{cn}[J]} \end{cases}$$

Optimizations

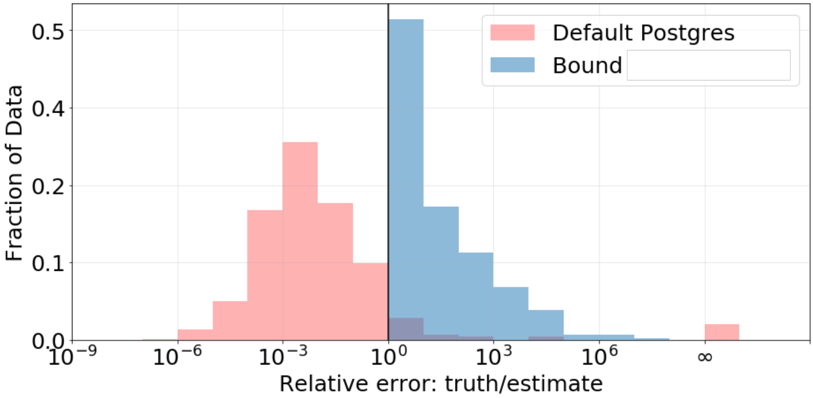
- Bound Formula Generation
- Partition Budgeting
  - Combats exponential runtime w.r.t. hash size
  - Non-monotonic behaviour
- Filter Predicates

Join Order Benchmark[1]

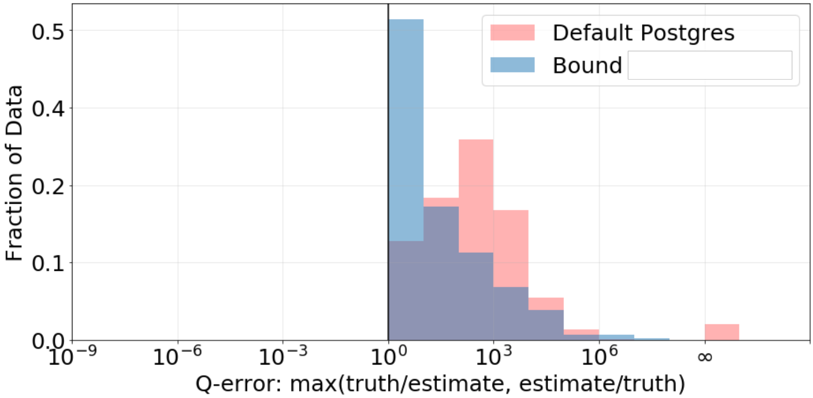
- Built on the IMDb dataset
  - 113 queries
  - 33 unique topoplogies
  - Skew!
  - Correlation!
  - Complex selection predicates!

[1] *How Good Are Query Optimizers, Really?* Leis et al. VLDB 2015.

Bound Relative Error Versus Postgres Relative Error



Bound Q-Error Versus Postgres Q-Error



Plan Execution Runtime (With Foreign Keys Indexes)

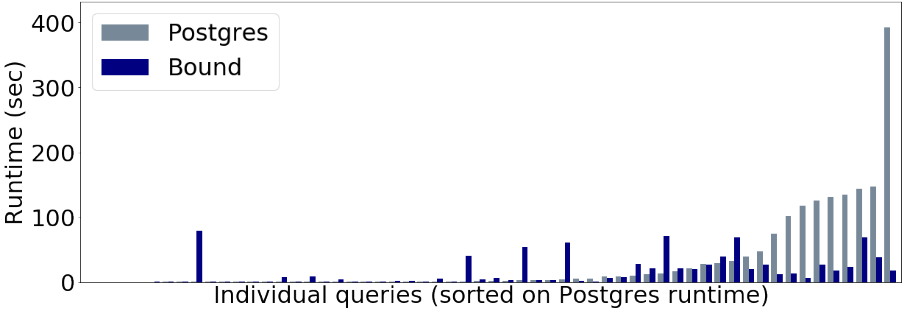


Figure: Linear scale runtime improvements over JOB queries.

- Total runtime:
  - Postgres: 3,190 seconds
  - Tightened Bound: 1,832 seconds

Plan Execution Runtime (With Foreign Keys Indexes)

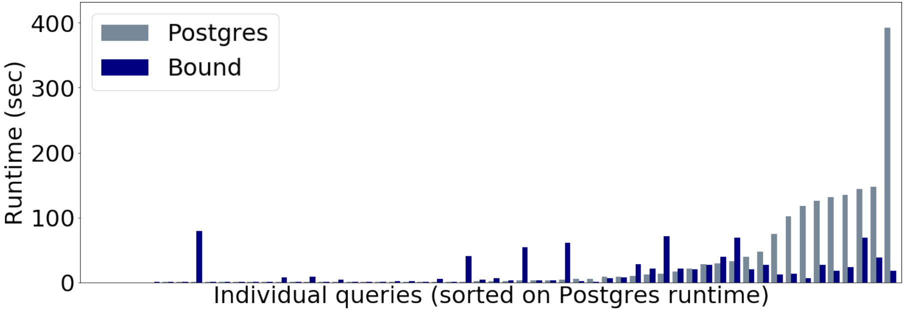


Figure: Linear scale runtime improvements over JOB queries.

- Total runtime:
  - Postgres: 3,190 seconds
  - Tightened Bound: 1,832 seconds

Plan Execution Runtime (No Foreign Key Indexes)

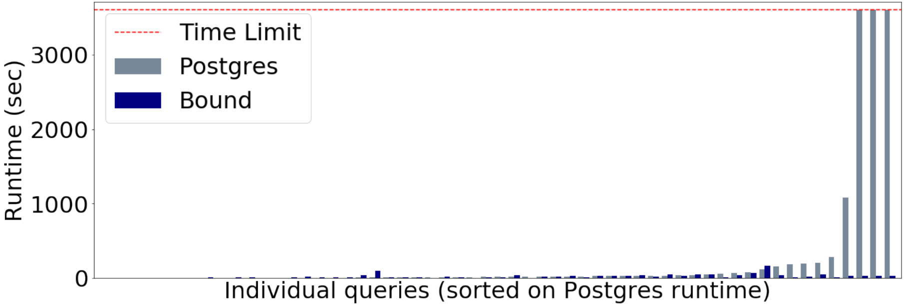


Figure: Linear scale plan execution time over JOB queries.

- Total runtime (including 1 hour cutoff for default Postgres):
  - Postgres: 21,125 seconds
  - Tightened Bound: 2,216 seconds

Plan Execution Runtime (No Foreign Key Indexes)

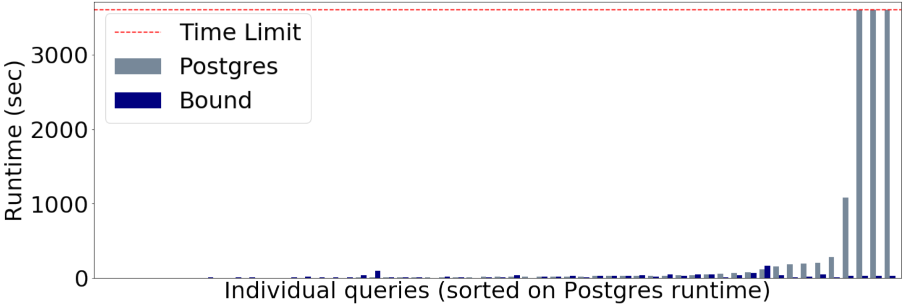


Figure: Linear scale plan execution time over JOB queries.

- Total runtime (including 1 hour cutoff for default Postgres):
  - Postgres: 21,125 seconds
  - Tightened Bound: 2,216 seconds

Takeaways

- Gains for very slow queries.
- On par for fast queries.
- Pessimistic but robust query optimization.

Acknowledgements

- Thank you to Jenny, Tomer, Laurel, Brandon, Jingjing, Tobin, Leilani, and Guna!
- This research is supported by NSF grant AITF 1535565 and III 1614738.

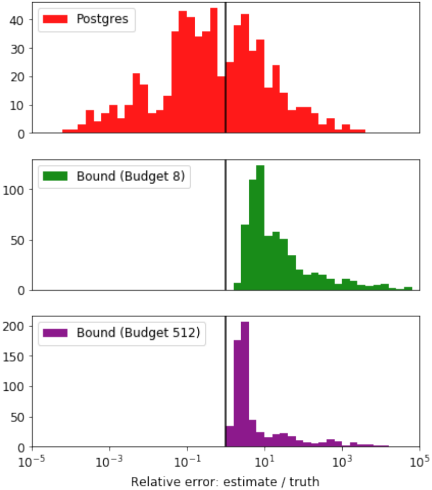
Googleplus Microbenchmark Examples

```
SELECT COUNT(*)
FROM
    community_44 AS t0,
    community_44 AS t1,
    community_44 AS t2,
    community_44 AS t3
WHERE
    t0.object = t1.subject AND
    t1.object = t2.subject AND
    t2.object = t3.subject AND
    t0.subject % 512 = 89 AND
    t3.object % 512 = 174;
```

Googleplus Microbenchmark Examples

```
SELECT COUNT(*)
FROM
    community_30 AS t0,
    community_30 AS t1,
    community_30 AS t2,
    community_30 AS t3,
    community_30 AS t4
WHERE
    t0.object = t1.subject AND
    t0.object = t2.subject AND
    t0.object = t3.subject AND
    t3.object = t4.subject AND
    t0.subject % 256 = 49 AND
    t1.object % 256 = 213 AND
    t2.object % 256 = 152 AND
    t4.object % 256 = 248;
    AND ci.movie_id = mc.movie_id;
```

Googleplus Progressive Bound Tightness



Example of Non-monotonic Behavior

$$Q(x, y, z, w) \text{ :- } R(z, y), S(y, z), T(z, w)$$

$$\underbrace{\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline 0 & 1 \\ \hline 1 & 0 \\ \hline 1 & 1 \\ \hline \end{array}}_{R} \bowtie \underbrace{\begin{array}{|c|c|}\hline y & z \\ \hline 0 & 0 \\ \hline 1 & 0 \\ \hline 2 & 1 \\ \hline 3 & 1 \\ \hline \end{array}}_{S} \bowtie \underbrace{\begin{array}{|c|c|}\hline z & w \\ \hline 0 & 0 \\ \hline 1 & 1 \\ \hline 2 & 2 \\ \hline 3 & 3 \\ \hline \end{array}}_{T} = \underbrace{\begin{array}{|c|c|c|c|}\hline x & y & z & w \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array}}_{Q}$$

Example of Non-monotonic Behavior

$$\underbrace{\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline 0 & 1 \\ \hline 1 & 0 \\ \hline 1 & 1 \\ \hline \end{array}}_{R} \bowtie \underbrace{\begin{array}{|c|c|}\hline y & z \\ \hline 0 & 0 \\ \hline 1 & 0 \\ \hline 2 & 1 \\ \hline 3 & 1 \\ \hline \end{array}}_{S} \bowtie \underbrace{\begin{array}{|c|c|}\hline z & w \\ \hline 0 & 0 \\ \hline 1 & 1 \\ \hline 2 & 2 \\ \hline 3 & 3 \\ \hline \end{array}}_{T} = \underbrace{\begin{array}{|c|c|c|c|}\hline x & y & z & w \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array}}_{Q}$$

$$|Q(x, y, z, w)| \leq \min \begin{cases} c_R \cdot d_S^y \cdot d_T^z \\ d_R^y \cdot c_S \cdot d_T^z \\ d_R^y \cdot d_S^z \cdot c_T \\ c_R \cdot c_T \end{cases}$$

$$c_{R(0)} \cdot d^y_{S(0,0)} \cdot d^z_{T(0)} = 4 \cdot 1 \cdot 1 = 4$$

Example of Non-monotonic Behavior

$$
\underbrace{\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline 0 & 1 \\ \hline 1 & 0 \\ \hline 1 & 1 \\ \hline \end{array}}_{R}
\bowtie
\underbrace{\begin{array}{|c|c|}\hline y & z \\ \hline 0 & 0 \\ \hline 1 & 0 \\ \hline 2 & 1 \\ \hline 3 & 1 \\ \hline \end{array}}_{S}
\bowtie
\underbrace{\begin{array}{|c|c|}\hline z & w \\ \hline 0 & 0 \\ \hline 1 & 1 \\ \hline 2 & 2 \\ \hline 3 & 3 \\ \hline \end{array}}_{T}
=
\underbrace{\begin{array}{|c|c|c|c|}\hline x & y & z & w \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array}}_{Q}
$$

$$
\left|Q(x,y,z,w)\right| \leq \min \begin{cases} c_R \cdot d_S^y \cdot d_T^z \\ d_R^y \cdot c_S \cdot d_T^z \\ d_R^y \cdot d_S^z \cdot c_T \\ c_R \cdot c_T \end{cases}
$$

$$
c_{R(0)} \cdot d^y_{S(0,0)} \cdot d^z_{T(0)} = 4 \cdot 1 \cdot 1 = 4
$$

Example of Non-monotonic Behavior

$$
\underbrace{\begin{array}{|c|c|}\hline x & y \\ \hline 0 & 0 \\ \hline 0 & 1 \\ \hline 1 & 0 \\ \hline 1 & 1 \\ \hline \end{array}}_{R}
\bowtie
\underbrace{\begin{array}{|c|c|}\hline y & z \\ \hline 0 & 0 \\ \hline 1 & 0 \\ \hline 2 & 1 \\ \hline 3 & 1 \\ \hline \end{array}}_{S}
\bowtie
\underbrace{\begin{array}{|c|c|}\hline z & w \\ \hline 0 & 0 \\ \hline 1 & 1 \\ \hline 2 & 2 \\ \hline 3 & 3 \\ \hline \end{array}}_{T}
=
\underbrace{\begin{array}{|c|c|c|c|}\hline x & y & z & w \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array}}_{Q}
$$

$$
\left|Q(x,y,z,w)\right| \leq \min \begin{cases} c_R \cdot d_S^y \cdot d_T^z \\ d_R^y \cdot c_S \cdot d_T^z \\ d_R^y \cdot d_S^z \cdot c_T \\ c_R \cdot c_T \end{cases}
$$

$$
c_{R(0)} \cdot d^y_{S(0,0)} \cdot d^z_{T(0)} = 4 \cdot 1 \cdot 1 = 4
$$

Example of Non-monotonic Behavior

- Define hash function $\text{hash}(u_i) = i\%2$.

$$\text{hash}(0) = \text{hash}(2) = 0$$
$$\text{hash}(1) = \text{hash}(3) = 1$$

Partitioned Relations

$$\text{hash}(y), \text{hash}(z) = \dots$$

$$0,0 \longrightarrow \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline\end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline\end{array}$$

$$0,1 \longrightarrow \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 2 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline\end{array} = \emptyset$$

$$1,0 \longrightarrow \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline\end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline\end{array}$$

$$1,1 \longrightarrow \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 3 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline\end{array} = \emptyset$$

$$0,0 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline\end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline\end{array}$$

$$0,1 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 2 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline\end{array} = \emptyset$$

$$1,0 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 0 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline\end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline\end{array}$$

$$1,1 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 3 & 1 \\ \hline\end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline\end{array} = \emptyset$$

$$\sum_{\substack{i,j \\ \in\{0,1\}}} \min \begin{cases} c_{R^{(i)}} \cdot d^{y}_{S^{(i,j)}} \cdot d^{z}_{T^{(j)}} \\ d^{y}_{R^{(i)}} \cdot c_{S^{(i,j)}} \cdot d^{z}_{T^{(j)}} \\ d^{y}_{R^{(i)}} \cdot d^{z}_{S^{(i,j)}} \cdot c_{T^{(j)}} \\ c_{R^{(i)}} \cdot c_{T^{(j)}} \end{cases} = \sum_{\substack{i,j \\ \in\{0,1\}}} \min \begin{cases} 2 \cdot 1 \cdot 1 \\ 2 \cdot 1 \cdot 1 \\ 2 \cdot 1 \cdot 2 \\ 2 \cdot 2 \end{cases} = \sum_{\substack{i,j \\ \in\{0,1\}}} 2 = 8$$

$$0,0 \longrightarrow \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline \end{array}$$

$$0,1 \longrightarrow \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 2 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline \end{array} = \emptyset$$

$$1,0 \longrightarrow \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array}$$

$$1,1 \longrightarrow \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 3 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline \end{array} = \emptyset$$

$$\sum_{\substack{i,j \\ \in\{0,1\}}} \min \begin{cases} c_{R^{(i)}} \cdot d^y_{S^{(i,j)}} \cdot d^z_{T^{(j)}} \\ d^y_{R^{(i)}} \cdot c_{S^{(i,j)}} \cdot d^z_{T^{(j)}} \\ d^y_{R^{(i)}} \cdot d^z_{S^{(i,j)}} \cdot c_{T^{(j)}} \\ c_{R^{(i)}} \cdot c_{T^{(j)}} \end{cases} = \sum_{\substack{i,j \\ \in\{0,1\}}} \min \begin{cases} 2 \cdot 1 \cdot 1 \\ 2 \cdot 1 \cdot 1 \\ 2 \cdot 1 \cdot 2 \\ 2 \cdot 2 \end{cases} = \sum_{\substack{i,j \\ \in\{0,1\}}} 2 = 8$$

$$0,0 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ \hline \end{array}$$

$$0,1 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 2 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline \end{array} = \emptyset$$

$$1,0 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 0 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 2 \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline 0 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array}$$

$$1,1 \longrightarrow \quad \begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 3 & 1 \\ \hline \end{array} \bowtie \begin{array}{|c|c|}\hline 1 & 1 \\ \hline 3 & 3 \\ \hline \end{array} = \emptyset$$

$$\sum_{\substack{i,j \\ \in\{0,1\}}} \min \begin{cases} c_{R^{(i)}} \cdot d^{y}_{S^{(i,j)}} \cdot d^{z}_{T^{(j)}} \\ d^{y}_{R^{(i)}} \cdot c_{S^{(i,j)}} \cdot d^{z}_{T^{(j)}} \\ d^{y}_{R^{(i)}} \cdot d^{z}_{S^{(i,j)}} \cdot c_{T^{(j)}} \\ c_{R^{(i)}} \cdot c_{T^{(j)}} \end{cases} = \sum_{\substack{i,j \\ \in\{0,1\}}} \min \begin{cases} 2 \cdot 1 \cdot 1 \\ 2 \cdot 1 \cdot 1 \\ 2 \cdot 1 \cdot 2 \\ 2 \cdot 2 \end{cases} = \sum_{\substack{i,j \\ \in\{0,1\}}} 2 = 8$$

Exponential Growth

- Sketch size (number of buckets) exponential in hash size.
  - Exponent = number of attributes in relation.
- Number of elements to sum up exponential in hash size.
  - Exponent = number of attributes in entire query.
- Non-monotonic bound behavior

Tuning Bucket Allocation

- Larger hash size $\implies$ more information $\implies$ tighter bounds, right?
  - Partitioning *unconditionally* covered attributes: yes.
  - Partitioning *conditionally* covered attributes: no.
  - Non-monotonic tradeoff space.

Non-Linearity of Degree Statistic

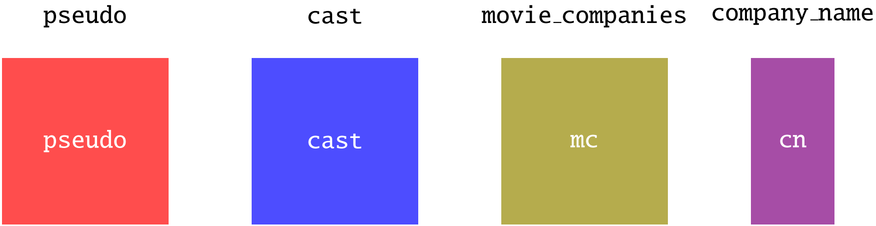
- Count is linear with respect to disjoint union!

$$\text{count}(A) + \text{count}(B) = \text{count}(A \cup B)$$

- Degree is not...

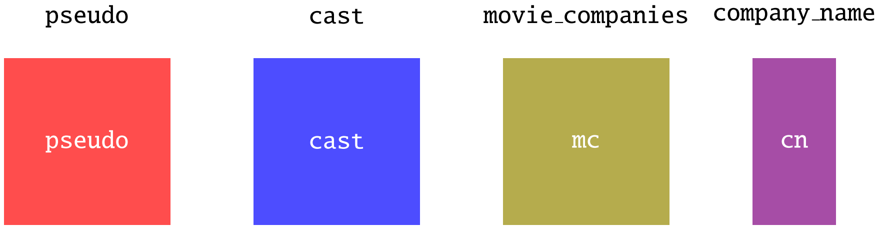
$$\text{degree}(A) + \text{degree}(B) \geq \text{degree}(A \cup B)$$

$$Q(x, y, z, w) :- pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$



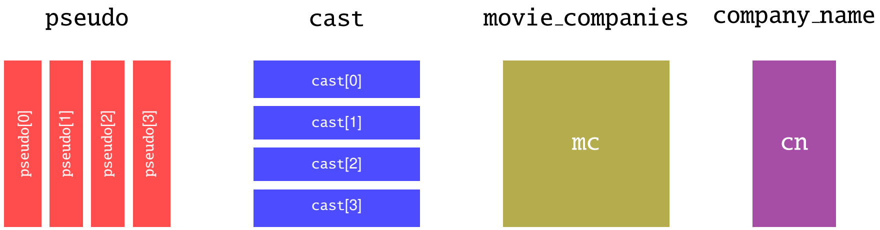
$$Q(x, y, z, w) :\!-\; pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$

$$c_{\texttt{pseudo}} \cdot d^{y}_{\texttt{cast}} \cdot d^{z}_{\texttt{mc}}$$



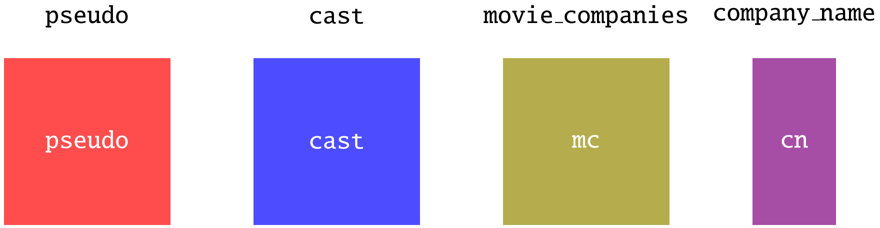
$$Q(x,y,z,w) :\!- \; pseudo(x,y), cast(y,z), mc(z,w), cn(w)$$

$$c_{\texttt{pseudo}} \cdot d^{y}_{\texttt{cast}} \cdot d^{z}_{\texttt{mc}}$$



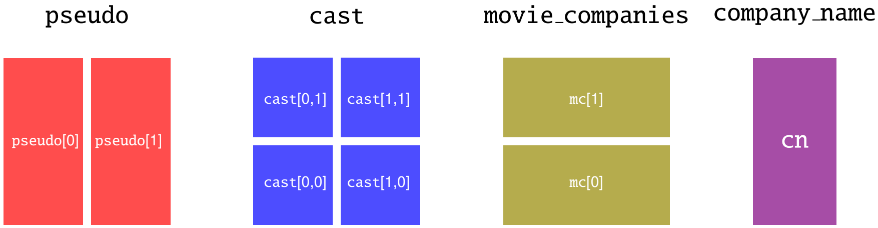
$$Q(x, y, z, w) :- pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$

$$d^{y}_{\texttt{pseudo}} \cdot c_{\texttt{cast}} \cdot d^{z}_{\texttt{mc}}$$



$$Q(x, y, z, w) :- pseudo(x, y), cast(y, z), mc(z, w), cn(w)$$

$$d^{y}_{\texttt{pseudo}} \cdot c_{\texttt{cast}} \cdot d^{z}_{\texttt{mc}}$$



Reformulated Bound Formula

- Old:

$$\left|Q(D)\right| \leq \sum_{\substack{J \in \\ \text{partition indexes}}} \left(\min_{\substack{b \in \\ \text{bounding formulas}}} b(Q(D[J])) \right)$$

- New:

$$\left|Q(D)\right| \leq \min_{\substack{b \in \\ \text{bounding formulas}}} \left(\sum_{\substack{J \in \\ \text{partition indexes} \\ \text{w.r.t. } b}} b(Q(D[J])) \right)$$