Announcements

• Project Milestone due on Friday

• Homework 4 posted; due next Friday

• There will be a short Homework 5, on transactions
Quick Recap

• Name 3 join processing algorithms
Outline

Algorithms for multi-joins

• AGM formula for maximum output size

• Generic-join algorithm matching that formula
Multi-join

- select * from R, S, T, … where …

- **Standard approach:**
  - Compute one join at a time
  - Optimizer chooses an “optimal” join order

- **Issues:**
  - Cardinality estimation is hard
  - Even “optimal” plan may be suboptimal
Plans Are Suboptimal

Because intermediate results are much larger than the final query answer
Example

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \land T(Z,X) \]

\[
\begin{align*}
\text{select} & \quad \ast & \quad \text{-- natural join} \\
\text{from} & \quad R, S, T \\
\text{where} & \quad R.Y = S.Y \text{ and } S.Z = T.Z \text{ and } T.X = R.X
\end{align*}
\]

Query plan

\[
\begin{array}{c}
\text{R}(X,Y) \\
\text{T}(Z,X) \\
\text{S}(Y,Z)
\end{array}
\]
Example

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \land T(Z,X) \]

```
select *  -- natural join
from R, S, T
where R.Y = S.Y and S.Z = T.Z and T.X = R.X
```

Query plan

\[
\begin{array}{c|c}
X & Y \\
0 & 1 \\
0 & 2 \\
0 & 3 \\
... & ...
\end{array}
\]

\[
\begin{array}{c|c}
X & Y \\
0 & N/2 \\
1 & 0 \\
2 & 0 \\
... & ...
\end{array}
\]
Example

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \land T(Z,X) \]

```
select *  -- natural join
from R, S, T
where R.Y = S.Y and S.Z = T.Z and T.X = R.X
```

Query plan

<table>
<thead>
<tr>
<th></th>
<th>R:</th>
<th>S: (same as R)</th>
<th>T: (same as R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0</td>
<td>N/2</td>
<td>0</td>
<td>N/2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>N/2</td>
<td>0</td>
<td>N/2</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ Ngo'2013 \]
Example

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \land T(Z,X) \]

```
select * -- natural join
from R, S, T
where R.Y = S.Y and S.Z = T.Z and T.X = R.X
```
Optimal Algorithm

To define “optimal” we need to answer two questions:

Q1: How large is the output of a query?

Q2: How can we compute it in time no larger than the largest output?
Worst-Case Optimality

Fix input statistics for $D$

• Runtime = $O(\max_{D \text{ satisfies stats}}(|Q(D)|))$

E.g. $R(X,Y) \Join S(Y,Z)$, $|R|, |S| \leq N$

• No other info: $|Output| \leq N$

• $S.Y$ is a key: $|Output| \leq N$

• $S.Y$ has degree $\leq d$: $|Output| \leq d \times N$

E.g. $R(X,Y) \Join S(Y,Z) \Join T(Z,X)$

$|Output| \leq N^{3/2}$
Worst-Case Optimality

Fix input statistics for $D$
• Runtime $= O(\max_{D \text{ satisfies stats}}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z), \ |R|, |S| \leq N$
Worst-Case Optimality

Fix input statistics for $D$

• Runtime $= O(\max_{D \text{ satisfies stats}}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z), \ |R|, |S| \leq N$

• No other info: $|Q(D)| \leq N^2$
Worst-Case Optimality

Fix input statistics for $D$

- Runtime = $O(\max_{D \text{ satisfies stats}}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z)$, $|R|, |S| \leq N$

- No other info: $|Q(D)| \leq N^2$

- $S.Y$ is a key:
Worst-Case Optimality

Fix input statistics for D
• Runtime = $O(\max_D \text{satisfies stats}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z)$, $|R|, |S| \leq N$
• No other info: $|Q(D)| \leq N^2$
• $S.Y$ is a key: $|Q(D)| \leq N$
Worst-Case Optimality

Fix input statistics for $D$

- Runtime = $O(\max_{D \text{ satisfies stats}}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z), \quad |R|, |S| \leq N$

- No other info: $|Q(D)| \leq N^2$
- $S.Y$ is a key: $|Q(D)| \leq N$
- $S.Y$ has degree $\leq d$: $|Output| \leq d \times N$
Worst-Case Optimality

Fix input statistics for $D$

- Runtime $= O(\max_D \text{satisfies stats}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z), \quad |R|, |S| \leq N$

- No other info: $|Q(D)| \leq N^2$
- $S.Y$ is a key: $|Q(D)| \leq N$
- $S.Y$ has degree $\leq d$: $|Q(D)| \leq d \times N$
Worst-Case Optimality

Fix input statistics for $D$

- Runtime = $O(\max_{D \text{ satisfies stats}}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z)$, $|R|, |S| \leq N$

- No other info: $|Q(D)| \leq N^2$

- S.Y is a key: $|Q(D)| \leq N$

- S.Y has degree $\leq d$: $|Q(D)| \leq d \times N$

E.g. $R(X,Y) \land S(Y,Z) \land T(Z,X)$
Worst-Case Optimality

Fix input statistics for $D$
- Runtime = $O(\max_{D \text{satisfies stats}}(|Q(D)|))$

E.g. $R(X,Y) \land S(Y,Z)$, $|R|, |S| \leq N$
- No other info: $|Q(D)| \leq N^2$
- $S.Y$ is a key: $|Q(D)| \leq N$
- $S.Y$ has degree $\leq d$: $|Q(D)| \leq d \times N$

E.g. $R(X,Y) \land S(Y,Z) \land T(Z,X)$
- No other info: $|Q(D)| \leq N^{3/2}$
The Two Questions

Q1: Given statistics, what is \( \max(|Q(D)|) \)?

Q2: How can we compute \( Q \) in time \( O(\max(|Q(D)|)) \)?
Simple Fact #1

• Consider any query:

\[ Q(X_1, \ldots, X_k) = R_1(\text{Vars}_1) \land \ldots \land R_m(\text{Vars}_m) \]

• Its output size is no larger than the product of all cardinalities:

\[ |Q| \leq |R_1| \times \ldots \times |R_m| \]
Graphs and Hypergraphs

• An undirected graph $G = (V, E)$ where each edge $e \in E$ is a set of two nodes
Graphs and Hypergraphs

- An undirected graph $G = (V, E)$ where each edge $e \in E$ is a set of two nodes.

- A hypergraph is $G = (V, E)$ where each edge is some set (of 1 or 2 or >2 nodes).
Conjunctive Queries are Hypergraphs

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) \]

\[ Q(x,y,z) = A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u) \]
Edge Cover

- An *edge cover* of a (hyper)graph is a subset of edges that contain all the vertices.
Edge Cover

• An *edge cover* of a (hyper)graph is a subset of edges that contain all the vertices
Edge Cover

- An *edge cover* of a (hyper)graph is a subset of edges that contain all the vertices.
Edge Cover

• An *edge cover* of a (hyper)graph is a subset of edges that contain all the vertices.
Simple Fact #2

• Consider any query:

\[ Q(X_1, ..., X_k) = R_1(Vars_1) \land ... \land R_m(Vars_m) \]

• Let \( R_{i_1}, R_{i_2}, ..., R_{i_n} \) be an edge cover. Then the output size is no larger than their product:

\[ |Q| \leq |R_{i_1}| \times \cdots \times |R_{i_n}| \]

Why?
Examples

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) \]
Examples

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) \]

- Edge covers:
  \[ R(x,y) \land S(y,z) \] or \[ R(x,y) \land T(z,x) \] or \[ S(y,z) \land T(z,x) \]
Examples

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) \]

- Edge covers:
  \[ R(x,y) \land S(y,z) \text{ or } R(x,y) \land T(z,x) \text{ or } S(y,z) \land T(z,x) \]

\[ |Q| \leq \min(|R| \times |S|, |R| \times |T|, |S| \times |T|) \]
Examples

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) \]

- Edge covers:
  \[ R(x,y) \land S(y,z) \text{ or } R(x,y) \land T(z,x) \text{ or } S(y,z) \land T(z,x) \]

\[ |Q| \leq \min(|R| \times |S|, |R| \times |T|, |S| \times |T|) \]

\[ Q(x,y,z,u) = A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u) \]
Examples

Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x)

• Edge covers:
  R(x,y) \land S(y,z) \text{ or } R(x,y) \land T(z,x) \text{ or } S(y,z) \land T(z,x)

\[ |Q| \leq \min(|R| \times |S|, |R| \times |T|, |S| \times |T|) \]

Q(x,y,z,u) = A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)

• Edge covers:
  A(x,y,z) \land B(x,y,u) \text{ or } A(x,y,z) \land C(x,z,u) \text{ or } …
Examples

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) \]

- Edge covers:
  \[ R(x,y) \land S(y,z) \text{ or } R(x,y) \land T(z,x) \text{ or } S(y,z) \land T(z,x) \]

\[ |Q| \leq \min(|R| \times |S|, |R| \times |T|, |S| \times |T|) \]

\[ Q(x,y,z,u) = A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u) \]

- Edge covers:
  \[ A(x,y,z) \land B(x,y,u) \text{ or } A(x,y,z) \land C(x,z,u) \text{ or } \ldots \]

\[ |Q| \leq \min(|A| \times |B|, |A| \times |C|, \ldots) \]
Fractional Edge Cover

- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers $w_e$, one for each edge $e$, such that, for every vertex $v$: $\sum_{e: v \in e} w_e \geq 1$
Fractional Edge Cover

- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers $w_e$, one for each edge $e$, such that, for every vertex $v$: $\sum_{e: v \in e} w_e \geq 1$
Fractional Edge Cover

- A **fractional edge cover** of a (hyper)graph is a set of non-negative numbers $w_e$, one for each edge $e$, such that, for every vertex $v$: $\sum_{e: v \in e} w_e \geq 1$
Fractional Edge Cover

- A fractional edge cover of a (hyper)graph is a set of non-negative numbers $w_e$, one for each edge $e$, such that, for every vertex $v$: $\sum_{e: v \in e} w_e \geq 1$
Fractional Edge Cover

- A fractional edge cover of a (hyper)graph is a set of non-negative numbers $w_e$, one for each edge $e$, such that, for every vertex $v$: $\sum_{e: v \in e} w_e \geq 1$.
Fractional Edge Cover

• A fractional edge cover of a (hyper)graph is a set of non-negative numbers $w_e$, one for each edge $e$, such that, for every vertex $v$: $\sum_{e: v \in e} w_e \geq 1$

• Fact: every edge cover is also a fractional edge cover. Why?
Not so Simple Fact #3

• Consider any query:

\[ Q(X_1, ..., X_k) = R_1(Vars_1) \land ... \land R_m(Vars_m) \]

• Let \( w_1, w_2, ..., w_m \) be a fractional edge cover. Then the output size is no larger than:

\[ |Q| \leq |R_1|^{w_1} \times ... \times |R_m|^{w_m} \]
### Examples

| Query       | $w_1, w_2, \ldots, w_m$ | $|R_1|^{w_1} \times \ldots \times |R_m|^{w_m}$ | $|Q| \leq \cdots$ |
|-------------|-------------------------|---------------------------------------------|-----------------|
| $R(x,y) \land S(y,z)$ | 1,1                     | $|R| \times |S|$                               | $\leq |R| \times |S|$ |


### Examples

| Query                          | \( w_1, w_2, \ldots, w_m \) | \( |R_1|^{w_1} \times \ldots \times |R_m|^{w_m} \) | \( |Q| \leq \ldots \) |
|-------------------------------|-----------------------------|---------------------------------|-------------------|
| \( R(x,y) \land S(y,z) \)    | 1, 1                        | \( |R| \times |S| \)               | \( \leq |R| \times |S| \) |
| \( R(x,y) \land S(y,z) \land T(z,x) \) |                           |                                 |                   |

\( \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \) for \( |A| \times |B| \times |C| \times |D| \)
## Examples

| Query                                      | $w_1, w_2, \ldots, w_m$ | $|R_1|^{w_1} \times \cdots \times |R_m|^{w_m}$ | $|Q| \leq \ldots$ |
|-------------------------------------------|--------------------------|-----------------------------------------------|-----------------|
| $R(x,y) \land S(y,z)$                     | 1, 1                     | $|R| \times |S|$                               | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$       | $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ | $(|R| \times |S| \times |T|)^{\frac{1}{2}}$ |                 |
## Examples

| Query                                      | $w_1, w_2, \ldots, w_m$ | $|R_1|^{w_1} \times \ldots \times |R_m|^{w_m}$ | $|Q| \leq \ldots$ |
|--------------------------------------------|--------------------------|-----------------------------------------------|-----------------|
| $R(x,y) \land S(y,z)$                      | 1,1                      | $|R| \times |S|$                              | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$        | $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ | $(|R| \times |S| \times |T|)^{\frac{1}{2}}$ | $\leq \min((|R| \times |S| \times |T|)^{\frac{1}{2}},\)$ |
|                                            | 1,1,0                    | $|R| \times |S|$                              | $|R| \times |S|, |R| \times |T|,$ |
|                                            | 1,0,1                    | $|R| \times |T|$                              | $|S| \times |T|)$ |
|                                            | 0,1,1                    | $|S| \times |T|$                              |                 |
# Examples

| Query                                      | \(w_1, w_2, \ldots, w_m\) | \(|R_1|^{w_1} \times \cdots \times |R_m|^{w_m}\) | \(|Q| \leq \ldots\) |
|-------------------------------------------|--------------------------|---------------------------------|---------------------|
| \(R(x,y) \land S(y,z)\)                | 1,1                      | \(|R| \times |S|\)                   | \(|R| \times |S|\)                   |
| \(R(x,y) \land S(y,z) \land T(z,x)\)   | \(1/2, 1/2, 1/2\)       | \((|R| \times |S| \times |T|)^{1/2}\) | \(\leq \min((|R| \times |S| \times |T|)^{1/2}, |R| \times |S|, |R| \times |T|, |S| \times |T|))\) |
| \(A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)\) | | | |
| | | | |
| | | | |
| | | | |
## Examples

| Query | $w_1$, $w_2$, ..., $w_m$ | $|R_1|^{w_1} \times ... \times |R_m|^{w_m}$ | $|Q| \leq ...$ |
|-------|------------------------|-----------------------------------|------------------|
| $R(x,y) \land S(y,z)$ | 1,1 | $|R| \times |S|$ | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$ | $1/2$, $1/2$, $1/2$ | $(|R| \times |S| \times |T|)^{1/2}$ | $\leq \min((|R| \times |S| \times |T|)^{1/2},$ $|R| \times |S|, |R| \times |T|, |S| \times |T|)$ |
| $A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)$ | $1/3$, $1/3$, $1/3$, $1/3$ | $(|A| \times |B| \times |C| \times |D|)^{1/3}$ | |

| Query | $w_1$, $w_2$, ..., $w_m$ | $|R_1|^{w_1} \times ... \times |R_m|^{w_m}$ | $|Q| \leq ...$ |
|-------|------------------------|-----------------------------------|------------------|
| $R(x,y) \land S(y,z)$ | 1,1 | $|R| \times |S|$ | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$ | $1/2$, $1/2$, $1/2$ | $(|R| \times |S| \times |T|)^{1/2}$ | $\leq \min((|R| \times |S| \times |T|)^{1/2},$ $|R| \times |S|, |R| \times |T|, |S| \times |T|)$ |
| $A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)$ | $1/3$, $1/3$, $1/3$, $1/3$ | $(|A| \times |B| \times |C| \times |D|)^{1/3}$ | |

| Query | $w_1$, $w_2$, ..., $w_m$ | $|R_1|^{w_1} \times ... \times |R_m|^{w_m}$ | $|Q| \leq ...$ |
|-------|------------------------|-----------------------------------|------------------|
| $R(x,y) \land S(y,z)$ | 1,1 | $|R| \times |S|$ | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$ | $1/2$, $1/2$, $1/2$ | $(|R| \times |S| \times |T|)^{1/2}$ | $\leq \min((|R| \times |S| \times |T|)^{1/2},$ $|R| \times |S|, |R| \times |T|, |S| \times |T|)$ |
| $A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)$ | $1/3$, $1/3$, $1/3$, $1/3$ | $(|A| \times |B| \times |C| \times |D|)^{1/3}$ | |

| Query | $w_1$, $w_2$, ..., $w_m$ | $|R_1|^{w_1} \times ... \times |R_m|^{w_m}$ | $|Q| \leq ...$ |
|-------|------------------------|-----------------------------------|------------------|
| $R(x,y) \land S(y,z)$ | 1,1 | $|R| \times |S|$ | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$ | $1/2$, $1/2$, $1/2$ | $(|R| \times |S| \times |T|)^{1/2}$ | $\leq \min((|R| \times |S| \times |T|)^{1/2},$ $|R| \times |S|, |R| \times |T|, |S| \times |T|)$ |
| $A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)$ | $1/3$, $1/3$, $1/3$, $1/3$ | $(|A| \times |B| \times |C| \times |D|)^{1/3}$ | |
## Examples

| Query | $w_1, w_2, \ldots, w_m$ | $|R_1|^{w_1} \times \ldots \times |R_m|^{w_m}$ | $|Q| \leq \ldots$ |
|-------|-------------------------|--------------------------------|-------------------|
| $R(x,y) \land S(y,z)$ | 1,1 | $|R| \times |S|$ | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$ | $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ | $(|R| \times |S| \times |T|)^{\frac{1}{2}}$ | $\leq \min((|R| \times |S| \times |T|)^{\frac{1}{2}}, |R| \times |S|, |R| \times |T|, |S| \times |T|)$ |
| $A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)$ | $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ | $(|A| \times |B| \times |C| \times |D|)^{\frac{1}{3}}$ | $\min(\ldots)$ |
## Examples

| Query                                      | $w_1$, $w_2$, ..., $w_m$ | $|R_1|^{w_1} \times \ldots \times |R_m|^{w_m}$ | $|Q| \leq \ldots$ |
|-------------------------------------------|--------------------------|-----------------------------------------------|------------------|
| $R(x,y) \land S(y,z)$                     | $1,1$                    | $|R| \times |S|$                                | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$       | $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ | $(|R| \times |S| \times |T|)^{\frac{1}{2}}$ | $\leq \text{min}((|R| \times |S| \times |T|)^{\frac{1}{2}},$ $|R| \times |S|, |R| \times |T|,$ $|S| \times |T|)$ |
| $A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)$ | $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ | $(|A| \times |B| \times |C| \times |D|)^{\frac{1}{3}}$ | $\text{min}(\ldots)$ |
| $R(x,y) \land S(y,z) \land T(z,u) \land K(u,v)$ |                           |                                              |                  |
## Examples

| Query | $w_1, w_2, \ldots, w_m$ | $|R_1|^{w_1} \times \ldots \times |R_m|^{w_m}$ | $|Q| \leq \ldots$ |
|-------|--------------------------|---------------------------------|-------------------|
| $R(x,y) \land S(y,z)$ | 1, 1 | $|R| \times |S|$ | $\leq |R| \times |S|$ |
| $R(x,y) \land S(y,z) \land T(z,x)$ | $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$ | $(|R| \times |S| \times |T|)^{\frac{1}{2}}$ | $\leq \min((|R| \times |S| \times |T|)^{\frac{1}{2}},\allowbreak |R| \times |S|, \allowbreak |R| \times |T|, \allowbreak |S| \times |T|)$ |
| $A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)$ | $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ | $(|A| \times |B| \times |C| \times |D|)^{\frac{1}{3}}$ | $\min(\ldots)$ |
| $R(x,y) \land S(y,z) \land T(z,u) \land K(u,v)$ | 1, 1, 1, 1 | $|R| \times |T| \times |K|$ | |
| | 1, 1, 1, 1 | $|R| \times |S| \times |K|$ | |
# Examples

| Query | \(w_1, w_2, \ldots, w_m\) | \(|R_1|^{w_1} \times \ldots \times |R_m|^{w_m}\) | \(|Q| \leq \ldots\) |
|-------|-----------------|-------------------------|------------------|
| \(R(x,y) \land S(y,z)\) | 1,1 | \(|R| \times |S|\) | \(\leq |R| \times |S|\) |
| \(R(x,y) \land S(y,z) \land T(z,x)\) | \(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\) | \((|R| \times |S| \times |T|)^{\frac{1}{2}}\) | \(\leq \min((|R| \times |S| \times |T|)^{\frac{1}{2}}, |R| \times |S|, |R| \times |T|, |S| \times |T|)\) |
| \(A(x,y,z) \land B(x,y,u) \land C(x,z,u) \land D(y,z,u)\) | \(1/3, 1/3, 1/3, 1/3\) | \((|A| \times |B| \times |C| \times |D|)^{1/3}\) | \(\min(\ldots)\) |
| \(R(x,y) \land S(y,z) \land T(z,u) \land K(u,v)\) | \(1,0,1,1\) | \(|R| \times |T| \times |K|\) |
| | \(1,1,0,1\) | \(|R| \times |S| \times |K|\) |
| | \(1, \frac{1}{2}, \frac{1}{2}, 1\) | (no need; why?) |
## Examples

| Query | \( w_1, w_2, \ldots, w_m \) | \(|R_1|^{w_1} \times \ldots \times |R_m|^{w_m}\) | \(|Q| \leq \ldots\) |
|-------|-----------------|-----------------|-----------------|
| \( R(x, y) \land S(y, z) \) | 1, 1 | \(|R| \times |S|\) | \(|R| \times |S|\) |
| \( R(x, y) \land S(y, z) \land T(z, x) \) | \( \frac{1}{2}, \frac{1}{2}, \frac{1}{2} \) | \((|R| \times |S| \times |T|)\)^{\frac{1}{2}}\) | \( \leq \min((|R| \times |S| \times |T|)^{\frac{1}{2}},\) \(|R| \times |S|,\) \(|R| \times |T|,\) \(|S| \times |T|)\) |
| \( A(x, y, z) \land B(x, y, u) \land C(x, z, u) \land D(y, z, u) \) | \( \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \) | \((|A| \times |B| \times |C| \times |D|)^{\frac{1}{3}}\) | \( \min(\ldots)\) |
| \( R(x, y) \land S(y, z) \land T(z, u) \land K(u, v) \) | 1, 0, 1, 1 | \(|R| \times |T| \times |K|\) | \( \min(|R| \times |T| \times |K|,\) \(|R| \times |S| \times |K|)\) |
| | 1, 1, 0, 1 | \(|R| \times |S| \times |K|\) | \( \) |
| | 1, \( \frac{1}{2}, \frac{1}{2}, 1 \) | (no need; why?) | \( \) |
Theorem $|Q| \leq \min_{w_1, \ldots, w_m} |R_1|^{w_1} \times \cdots \times |R_m|^{w_m}$

This is called the AGM bound* of $Q$. It is tight.

Note: it suffices to consider only those fractional edge covers $w_1, \ldots, w_m$ that are not convex combinations of others.

We will prove tightness on a special case.

But first, let’s discuss an algorithm for computing $Q$ with this runtime.

*Atserias, Grohe, Marx introduced this bound
Generic Join – Overview

• Choose a variable order

• Sort every relation $R_i$ according to this order: time is $O(|R_i| \log |R_i|) = \tilde{O}(|R_i|)$

• *Generic join* assumes relations are sorted; it computes $Q$ in time $\tilde{O}(\text{AGM}(Q))$

• “Worst case optimal”
Generic Join – The Intersection

*Intersection* is the main building block of G.J.

\[ Q(x) = R(x) \land S(x) \]

- Discuss merge-join in class – what is runtime?
Generic Join – The Intersection

Intersection is the main building block of G.J.

\[ Q(x) = R(x) \land S(x) \]

- Discuss merge-join in class – what is runtime?

- Edge covers of Q: 1,0 and 0,1; \(|Q| \leq \min(|R|, |S|)\)
Generic Join – The Intersection

*Intersection* is the main building block of G.J.

\[ Q(x) = R(x) \land S(x) \]

- Discuss merge-join in class – what is runtime?

- Edge covers of \( Q \): 1,0 and 0,1; \( |Q| \leq \min(|R|, |S|) \)
- Discuss improved merge-join in class

Runtime: \( \tilde{O}(\min(|R|, |S|)) \)
Generic Join Algorithm

Let $x$ be the first variable
Let $R_{i1}, R_{i2}, \ldots$ be all relations containing $x$
Compute $D = \Pi_x(R_{i1}) \cap \Pi_x(R_{i2}) \cap \ldots$
for every value $v \in D$ do:
  Compute $Q$,
  where $R_{i1}, R_{i2}, \ldots$ are restricted to $x = v$

needs to be done in time $\tilde{O}(\min_j \Pi_x(R_j))$
Generic Join Example

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) \]

Assume \(|R|=|S|=|T|=N\), then:

\[ A = \Pi_x(R(x,y)) \cap \Pi_x(T(z,x)) \]

\[ |Q| \leq N^{3/2} \]
Generic Join Example

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x), \]
Assume \(|R| = |S| = |T| = N\), then:

\[ A = \Pi_x(R(x,y)) \land \Pi_x(T(z,x)) \]
for a in A do
  /* compute \( Q(a,y,z) = R(a,y) \land S(y,z) \land T(z,a) \) */
Generic Join Example

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x), \]
Assume \(|R| = |S| = |T| = N\), then:

\[ A = \Pi_x(R(x,y)) \cap \Pi_x(T(z,x)) \]

\textbf{for} \ a \ \textbf{in} \ A \ \textbf{do}

/* compute \( Q(a,y,z) = R(a,y) \land S(y,z) \land T(z,a) \) */

\[ B = \Pi_y(R(a,y)) \cap \Pi_y(S(y,z)) \]

\textbf{for} \ b \ \textbf{in} \ B \ \textbf{do}

/* compute \( Q(a,b,z) = R(a,b) \land S(b,z) \land T(z,a) \) */

\( |Q| \leq N^{3/2} \)
Generic Join Example

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x), \]
Assume \(|R|=|S|=|T|=N\), then:

\[ |Q| \leq N^{3/2} \]

\[ A = \Pi_x(R(x,y)) \cap \Pi_x(T(z,x)) \]
\[ \text{for } a \text{ in } A \text{ do} \]
\[ /* \text{compute } Q(a,y,z) = R(a,y) \land S(y,z) \land T(z,a) */ \]
\[ B = \Pi_y(R(a,y)) \cap \Pi_y(S(y,z)) \]
\[ \text{for } b \text{ in } B \text{ do} \]
\[ /* \text{compute } Q(a,b,z) = R(a,b) \land S(b,z) \land T(z,a) */ \]
\[ C = \Pi_z(S(b,z)) \cap \Pi_z(T(z,a)) \]
\[ \text{for } c \text{ in } C \text{ do} \]
\[ \text{output } (a,b,c) \]
Generic Join Example

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x), \]
Assume \(|R|=|S|=|T|=N\), then:

\[ A = \Pi_x(R(x,y)) \cap \Pi_x(T(z,x)) \]

\textbf{for} a \textbf{in} A \textbf{do}

\textit{/* compute } Q(a,y,z) = R(a,y) \land S(y,z) \land T(z,a) */
\n\textbf{B} = \Pi_y(R(a,y)) \cap \Pi_y(S(y,z))

\textbf{for} b \textbf{in} B \textbf{do}

\textit{/* compute } Q(a,b,z) = R(a,b) \land S(b,z) \land T(z,a) */
\n\textbf{C} = \Pi_z(S(b,z)) \cap \Pi_z(T(z,a))

\textbf{for} c \textbf{in} C \textbf{do}

output (a,b,c)

\[ |Q| \leq N^{3/2} \]

\textbf{Runtime: } \tilde{O}(N^{3/2})
Discussion

• All relations need to be presorted, or indexed

• Runtime is guaranteed to be worst-case optimal, \textit{no matter} what variable order we choose

• In practice, the variable order \textit{does matter}; in class: discuss $R(x,y) \land S(y,z)$
Comparison to Naïve Nested Loop

Naïve nested loop:

// tuple at a time:
For t1 in R1 do
  for t2 in R2 do
    ...

Generic join:
A = ∩ domains for x
For x in A do
  B = ∩ domains for y
  For y in B do
    C = ∩ domains for z
    For z in C do
      ...

Naïve nested loop:

// tuple at a time:
For t1 in R1 do
  for t2 in R2 do
    ...

// value at a time:
For x in Domain do
  For y in Domain do
    For z in Domain do
      ...

Generic join
A = \cap domains for x
For x in A do
  B = \cap domains for y
  For y in B do
    C = \cap domains for z
    For z in C do
      ...

Comparison to Naïve Nested Loop

Naïve nested loop:

// tuple at a time:
For t1 in R1 do
  for t2 in R2 do
    ...

// value at a time:
For x in Domain do
  For y in Domain do
    For z in Domain do
      ...

Generic-join

A = ∩ domains for x
For x in A do
  B = ∩ domains for y
  For y in B do
    C = ∩ domains for z
    For z in C do
      ...
Tightness

• There exists instances $R_1, R_2, \ldots$ such that the size of the query’s output is $\text{AGM}(Q)$

• Proof is simple and instructive; we will show for special case $|R_1| = \ldots = |R_m| = N$

• In this case $\text{AGM}(Q) = N^{\min(w_1 + \ldots + w_m)}$
Fractional Edge Covering Number

- The fractional edge covering number of a hypergraph is $\rho^* = \min \sum_e w_e$, where the minimum is over all fractional edge covers of the hypergraph.

**Fact** Assume $|R_1| = \ldots = |R_m| = N$. Then $AGM(Q) = N\rho^*$
Fractional Vertex Packing

• A fractional vertex packing of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x: x \in e} v_x \leq 1$
A fractional vertex packing of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum x v_x \leq \sum e w_e$
A **fractional vertex packing** of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x: x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$
Fractional Vertex Packing

- A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$
Fractional Vertex Packing

- A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

$\rho^* = 1$
Fractional Vertex Packing

A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers \( v_x \), one for each node \( x \), such that, for every edge \( e \): 
\[
\sum_{x : x \in e} v_x \leq 1
\]

**Fact** For any \( v, w \): 
\[
\sum_x v_x \leq \sum_e w_e
\]

**Theorem** 
\[
\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e
\]
Fractional Vertex Packing

• A **fractional vertex packing** of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x: x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

\[ \rho^* = 1 \]
Fractional Vertex Packing

- A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

\[\rho^* = 3/2\]

\[\rho^* = 1\]
Fractional Vertex Packing

- A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x: x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$
**Fractional Vertex Packing**

- A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$
Fractional Vertex Packing

- A **fractional vertex packing** of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x : x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

\[ \rho^* = 1 \]

\[ \rho^* = 3/2 \]

\[ \rho^* = 3 \]
Fractional Vertex Packing

- A **fractional vertex packing** of a (hyper)graph is a set of non-negative numbers $v_x$, one for each node $x$, such that, for every edge $e$: $\sum_{x : x \in e} v_x \leq 1$

**Fact** For any $v, w$: $\sum_x v_x \leq \sum_e w_e$

**Theorem** $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

\[ \rho^* = \begin{cases} 1 & \text{for } \rho^* = 1 \\ \frac{3}{2} & \text{for } \rho^* = \frac{3}{2} \\ 3 & \text{for } \rho^* = 3 \end{cases} \]
The Bound is Tight

**Fact** Fix a fractional vertex packing $\mathbf{v} = (v_x)_{x \in \text{Nodes}}$. Then there exists a database such that $|R_1| \leq N$, ..., $|R_m| \leq N$ and $|Q| = N \sum_x v_x$. 
The Bound is Tight

**Fact** Fix a fractional vertex packing \( v = (v_x)_{x \in \text{Nodes}} \). Then there exists a database such that \( |R_1| \leq N, \ldots, |R_m| \leq N \) and \( |Q| = N \sum_{x} v_x \)

**Proof.** For every relation \( R_j \) with variables \( x_{i_1}, x_{i_2}, \ldots \) define the instance \( |R_j| = [N^{v_{i_1}}] \times [N^{v_{i_2}}] \times \ldots \) where \([k] = \{1,2,\ldots,k\}\).
The Bound is Tight

**Fact** Fix a fractional vertex packing \( v = (v_x)_{x \in \text{Nodes}} \). Then there exists a database such that \( |R_1| \leq N, \ldots, |R_m| \leq N \) and \( |Q| = N^{\sum v_x} \)

**Proof.** For every relation \( R_j \) with variables \( x_{i_1}, x_{i_2}, \ldots \) define the instance \( |R_j| = [N^{v_{i_1}}] \times [N^{v_{i_2}}] \times \ldots \) where \([k] = \{1,2,\ldots,k\}\). Then:

(\(a\)) \( |R_j| = N^{v_{i_1} + v_{i_2} + \ldots} \leq N \) (why?)
The Bound is Tight

**Fact** Fix a fractional vertex packing \( v = (v_x)_{x \in \text{Nodes}} \). Then there exists a database such that
\[ R_1| \leq N, \ldots, |R_m| \leq N \] and \( |Q| = N^{\sum_x v_x} \)

**Proof.** For every relation \( R_j \) with variables \( x_{i_1}, x_{i_2}, \ldots \)
define the instance \( |R_j| = [N^{v_{i_1}}] \times [N^{v_{i_2}}] \times \ldots \)
where \([k] = \{1,2,…,k\}\). Then:
(a) \( |R_j| = N^{v_{i_1} + v_{i_2} + \cdots} \leq N \) (why?)
(b) \( |Q| = N^{\sum_x v_x} \) (why?)
Example

$Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x)$,

- Assume $|R| = |S| = |T| = N$. 

Define $D_x = \lceil N^{1/2} \rceil = \{1, 2, \ldots, N^{1/2}\}$,

- $D_y = \lceil N^{1/2} \rceil$

- $D_z = \lceil N^{1/2} \rceil$

Then $|R| = |S| = |T| = N$, $Q = D_x \times D_y \times D_z$ and $|Q| = N^{3/2}$. 
Example

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x), \]

• Assume \(|R| = |S| = |T| = N\).

• Optimal vertex packing: \(v_x = v_y = v_z = 1/2\)
Example

\( Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x) , \)

• Assume \(|R|=|S|=|T|=N.\)

• Optimal vertex packing: \( v_x = v_y = v_z = 1/2 \)

• Define: \( D_x = [N^{1/2}] = \{1, 2, \ldots, N^{1/2}\} \)
  \( D_y = [N^{1/2}] \)
  \( D_z = [N^{1/2}] \)
Example

Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x),

• Assume |R| = |S| = |T| = N.

• Optimal vertex packing: \( v_x = v_y = v_z = 1/2 \)

• Define:
  \[
  D_x = \left\lfloor N^{1/2} \right\rfloor = \{1, 2, \ldots, N^{1/2}\}
  
  D_y = \left\lfloor N^{1/2} \right\rfloor
  
  D_z = \left\lfloor N^{1/2} \right\rfloor
  
  \]

• Define
  \[
  R = D_x \times D_y, \quad S = D_y \times D_z, \quad T = D_z \times D_x.
  \]
Example

\[ Q(x,y,z) = R(x,y) \land S(y,z) \land T(z,x), \]

- Assume \(|R| = |S| = |T| = N\).
- Optimal vertex packing: \(v_x = v_y = v_z = 1/2\)
- Define: \(D_x = [N^{1/2}] = \{1, 2, \ldots, N^{1/2}\}\)
  \(D_y = [N^{1/2}]\)
  \(D_z = [N^{1/2}]\)
- Define \(R = D_x \times D_y, S = D_y \times D_z, T = D_z \times D_x\).
- Then \(|R| = |S| = |T| = N\),
  \(Q = D_x \times D_y \times D_z\) and \(|Q| = N^{3/2}\)
Keys

\[ R(X,Y) \land S(Y,Z), \quad |R|, |S| \leq N \]

- No other info: \[ |Q(D)| \leq N^2 \]
- \( S.Y \) is a key: \[ |Q(D)| \leq N \]

The *Query Expansion* method:

- If \( Y \) is a key in some relation \( S \), then add all attributes of \( S \) relations containing \( Y \)
- Compute \( \text{AGM}(Q_{\text{expanded}}) \)
Examples

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \]
Examples

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \]

- \[ Q^{\text{exp}}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z), \]

- \[ \text{AGM}(Q^{\text{exp}}) = |R| \]

- \[ Q^{\text{exp}}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z) \land T(Z,X), \]

- \[ \text{AGM}(Q^{\text{exp}}) = \min(|R|, |S| \times |T|) \]
Examples

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \]

- \( Q^{\text{exp}}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z) \),
- Edge cover: 1,0
- \( \text{AGM}(Q^{\text{exp}}) = |R| \)
Examples

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \]

- \[ Q^{\exp}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z), \]
- Edge cover: 1,0
- \[ \text{AGM}(Q^{\exp}) = |R| \]

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \land T(Z,X) \]
Examples

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \]

- \[ Q^{\text{exp}}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z), \]
- Edge cover: 1,0
- \[ \text{AGM}(Q^{\text{exp}}) = |R| \]

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \land T(Z,X) \]

- \[ Q^{\text{exp}}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z) \land T(Z,X) \]
Examples

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \]
- \( Q^{\text{exp}}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z) \),
- Edge cover: 1,0
- \( \text{AGM}(Q^{\text{exp}}) = |R| \)

\[ Q(X,Y,Z) = R(X,Y) \land S(Y,Z) \land T(Z,X) \]
- \( Q^{\text{exp}}(X,Y,Z) = R(X,Y,Z) \land S(Y,Z) \land T(Z,X) \)
- Edge covers: 1,0,0 or 0,1,1
- \( \text{AGM}(Q^{\text{exp}}) = \min(|R|, |S| \times |T|) \)
Summary

Given cardinalities of all input tables:
• AGM bound gives upper bound on query size
• GJ computes the query in this time

Generic Join:
• A nested loop algorithm
• No longer one-join-at-a-time
• Theoretical optimality means it will be efficient for very expensive queries; less so for cheaper queries