CSE544
Data Management

Lectures 12
Advanced Query Processing

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

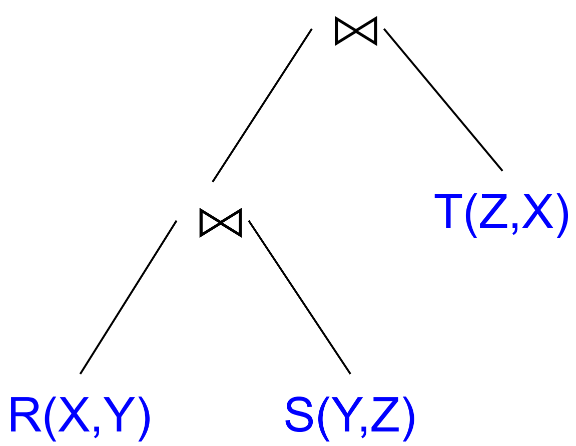
[Ngo'2013]

Example

Q(X,Y,Z) = R(X,Y) ∧ S(Y,Z) ∧ 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



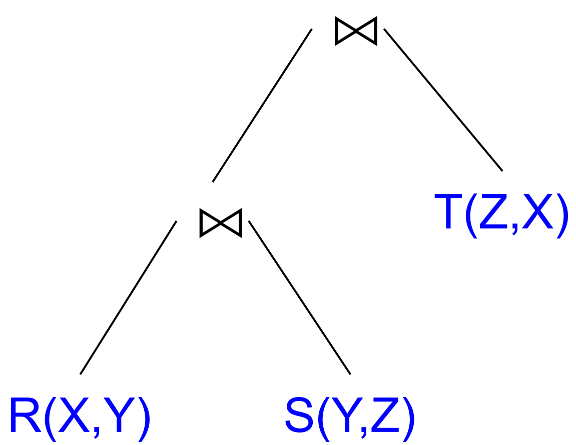
[Ngo'2013]

Example

$Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

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

Query plan



R: (N rows)

X	Y
0	1
0	2
0	3
...	...
0	N/2
1	0
2	0
...	...
N/2	0

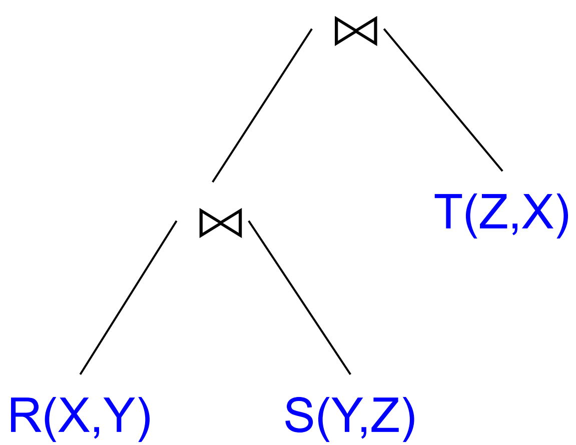
[Ngo'2013]

Example

$Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

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

Query plan



N

R:

X	Y
0	1
0	2
0	3
...	...
0	N/2
1	0
2	0
...	...
N/2	0

S: (same as R)

Y	Z
0	1
0	2
0	3
...	...
0	N/2
1	0
2	0
...	...
N/2	0

T: (same as R)

Z	X
0	1
0	2
0	3
...	...
0	N/2
1	0
2	0
...	...
N/2	0

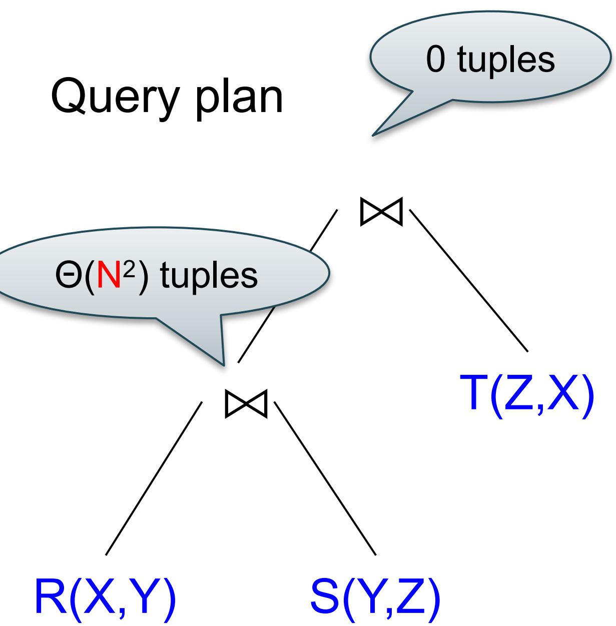
[Ngo'2013]

Example

$Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

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

Query plan



N (rows)

R:

X	Y
0	1
0	2
0	3
...	...
0	N/2
1	0
2	0
...	...
N/2	0

S: (same as R)

Y	Z
0	1
0	2
0	3
...	...
0	N/2
1	0
2	0
...	...
N/2	0

T: (same as R)

Z	X
0	1
0	2
0	3
...	...
0	N/2
1	0
2	0
...	...
N/2	0

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

Worst-Case Optimality

Fix input statistics for D

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

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

Worst-Case Optimality

Fix input statistics for D

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

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

- 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. $R(X,Y) \wedge S(Y,Z)$, $|R|, |S| \leq N$

- 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. $R(X,Y) \wedge S(Y,Z)$, $|R|, |S| \leq N$

- 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. $R(X,Y) \wedge S(Y,Z)$, $|R|, |S| \leq N$

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

Worst-Case Optimality

Fix input statistics for D

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

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

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

Worst-Case Optimality

Fix input statistics for D

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

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

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

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

Worst-Case Optimality

Fix input statistics for D

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

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

- 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. $R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

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:

$$Q(X_1, ..., X_k) = R_1(Vars_1) \wedge ... \wedge R_m(Vars_m)$$

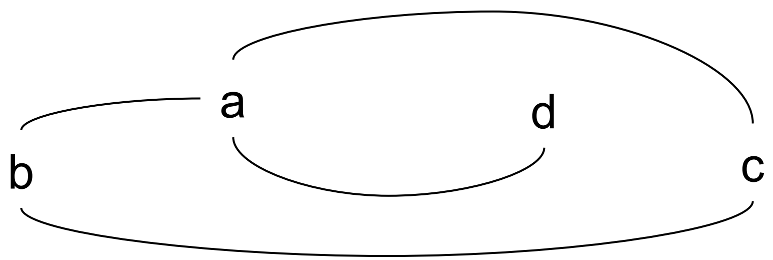
- Its output size is no larger than the product of all cardinalities:

$$|Q| \leq |R_1| \times ... \times |R_m|$$

Why?

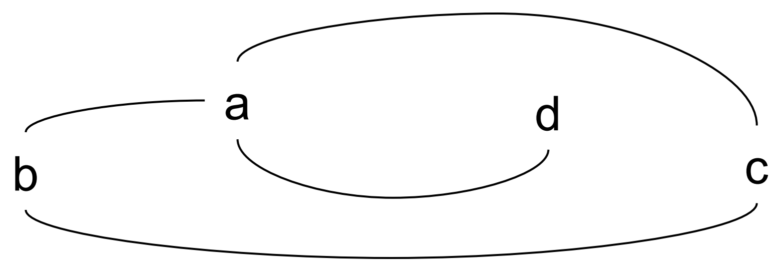
Graphs and Hypergraphs

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

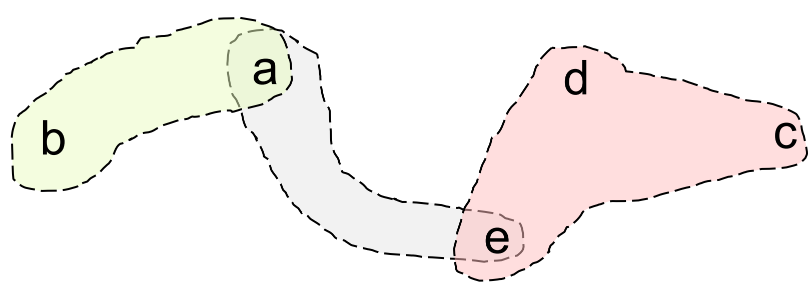


Graphs and Hypergraphs

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

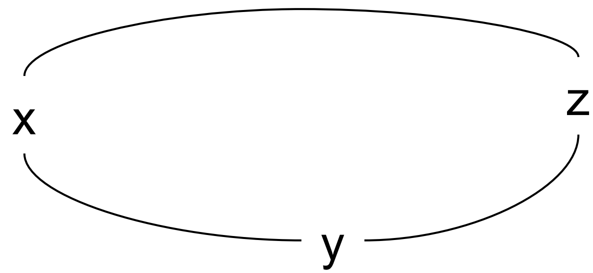


- A hypergraph is G = (V, E) where each edge is some set (of 1 or 2 or >2 nodes)

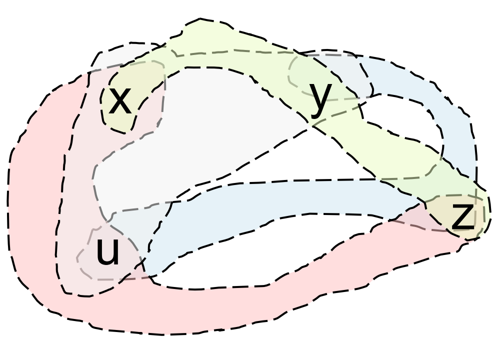


Conjunctive Queries are Hypergraphs

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

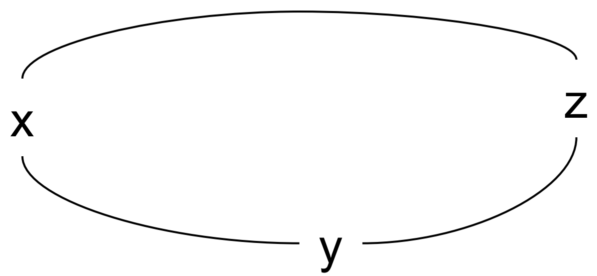


$Q(x,y,z) =$
$A(x,y,z) \wedge B(x,y,u) \wedge C(x,z,u) \wedge D(y,z,u)$



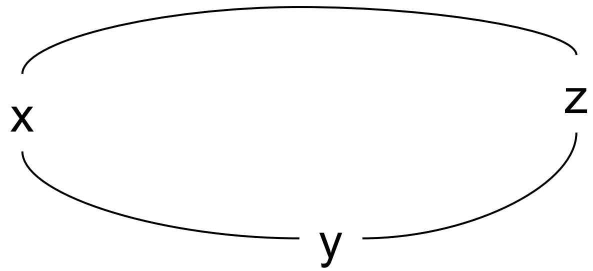
Edge Cover

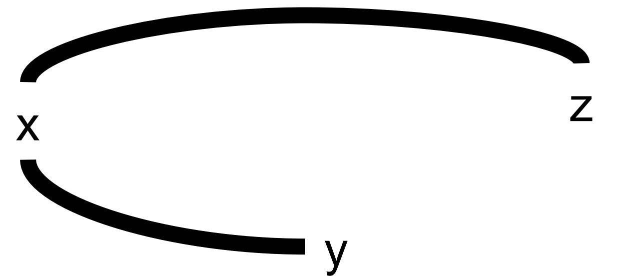
- An *edge cover* of a (hyper)graph is a subset of edges that contain all the vertices



Edge Cover

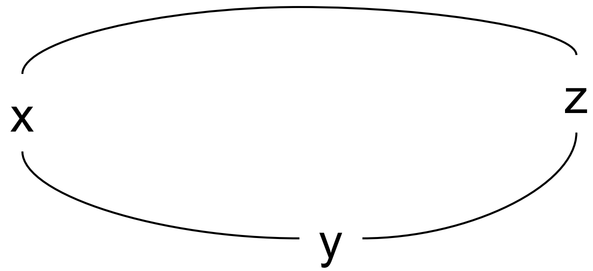
- An *edge cover* of a (hyper)graph is a subset of edges that contain all the vertices

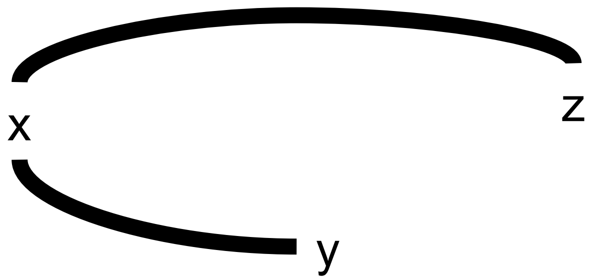


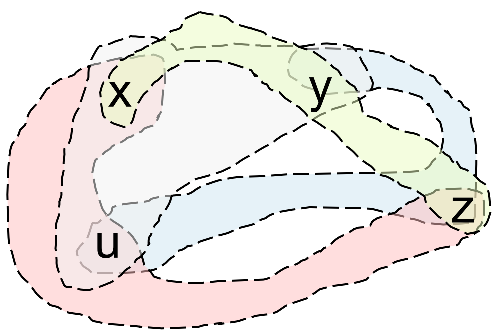


Edge Cover

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

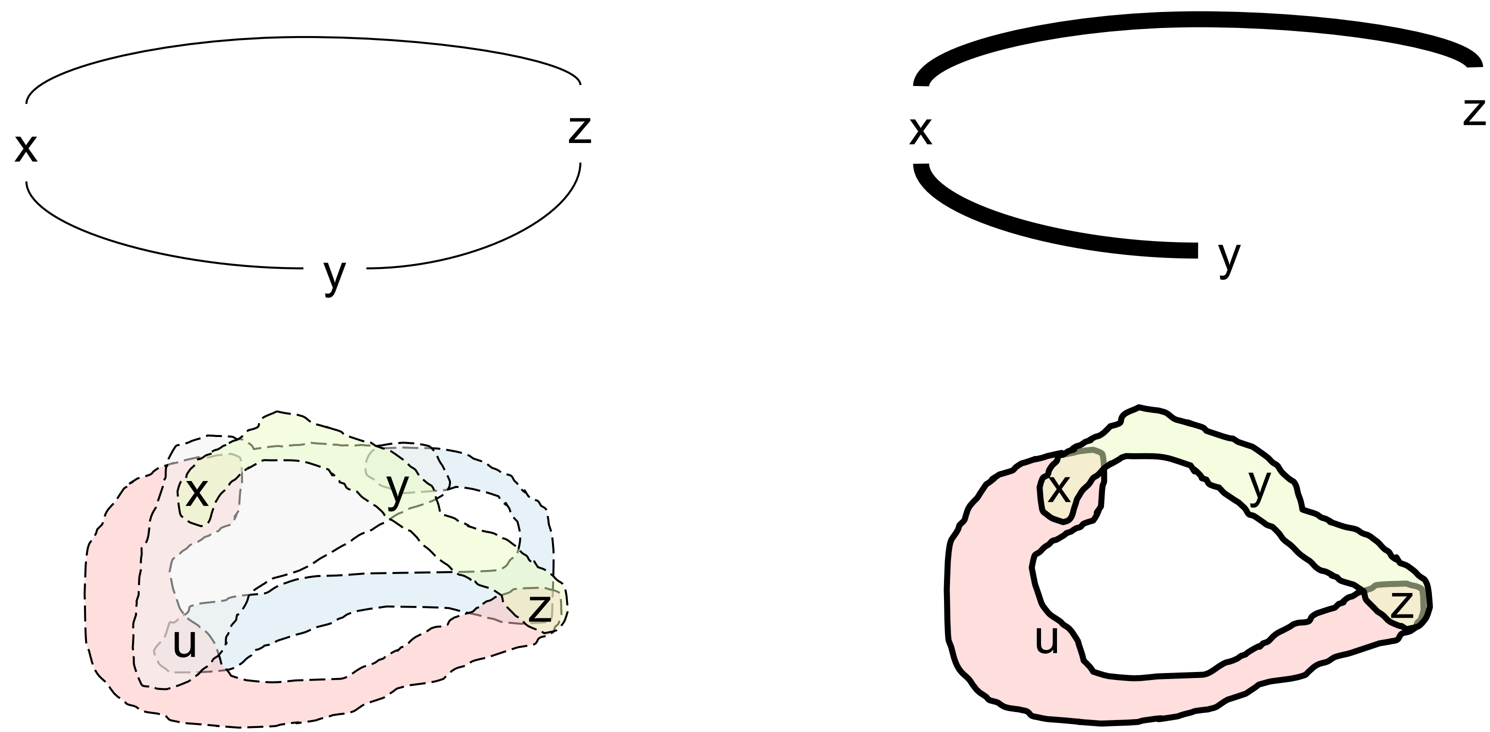






Edge Cover

- An <u>*edge cover*</u> of a (hyper)graph is a subset of edges that contain all the vertices



Simple Fact #2

- Consider any query:

$$Q(X_1, \ldots, X_k) = R_1(Vars_1) \wedge \ldots \wedge R_m(Vars_m)$$

- Let $R_{i_1}, R_{i_2}, \ldots, R_{i_n}$ 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

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

Examples

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

- Edge covers:
  $R(x,y) \wedge S(y,z)$ or $R(x,y) \wedge T(z,x)$ or $S(y,z) \wedge T(z,x)$

Examples

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

- Edge covers:
  $R(x,y) \wedge S(y,z)$ or $R(x,y) \wedge T(z,x)$ or $S(y,z) \wedge T(z,x)$

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

Examples

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

- Edge covers:
  $R(x,y) \wedge S(y,z)$ or $R(x,y) \wedge T(z,x)$ or $S(y,z) \wedge T(z,x)$

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

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

Examples

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

- Edge covers:
  $R(x,y) \wedge S(y,z)$ or $R(x,y) \wedge T(z,x)$ or $S(y,z) \wedge T(z,x)$

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

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

- Edge covers:
  $A(x,y,z) \wedge B(x,y,u)$ or $A(x,y,z) \wedge C(x,z,u)$ or …

Examples

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

- Edge covers:
  $R(x,y) \wedge S(y,z)$ or $R(x,y) \wedge T(z,x)$ or $S(y,z) \wedge T(z,x)$

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

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

- Edge covers:
  $A(x,y,z) \wedge B(x,y,u)$ or $A(x,y,z) \wedge C(x,z,u)$ or …

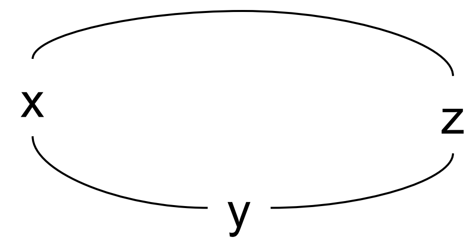
$|Q|\ | \leq \min(|A| \times |B|, |A| \times |C|, \ldots)$

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: $\sum_{e:\, v \in e} w_e \geq 1$

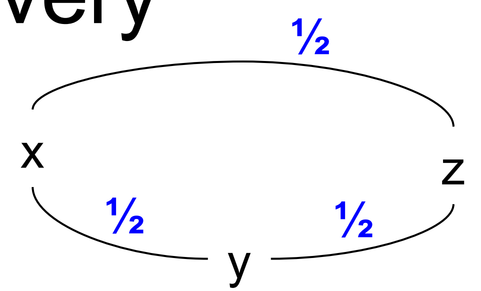
Fractional Edge Cover

- 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: $\sum_{e:\, v \in e} w_e \geq 1$



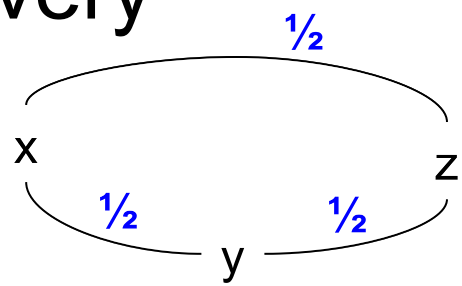
Fractional Edge Cover

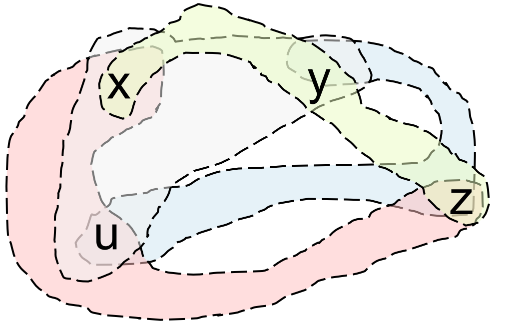
- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers w_e, one for each edge e, such that, for every vertex v: $\sum_{e:\, v \in e} w_e \geq 1$



Fractional Edge Cover

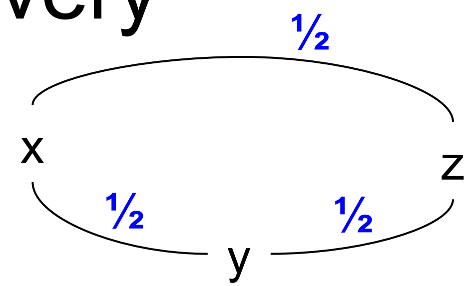
- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers w_e, one for each edge e, such that, for every vertex v: $\sum_{e:\, v \in e} w_e \geq 1$

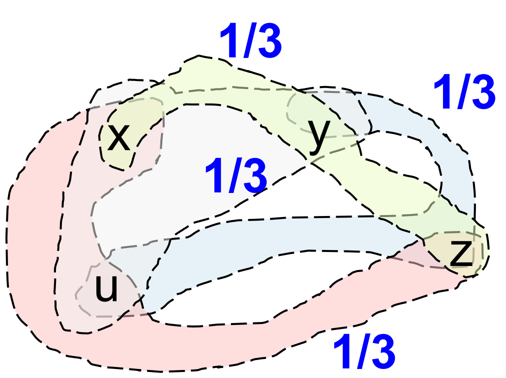




Fractional Edge Cover

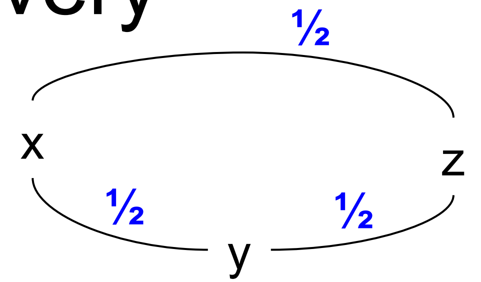
- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers w_e, one for each edge e, such that, for every vertex v: $\sum_{e:\, v \in e} w_e \geq 1$



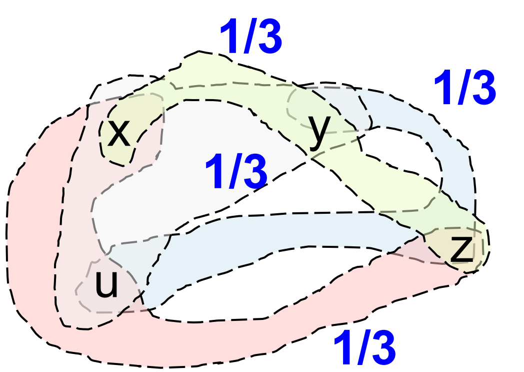


Fractional Edge Cover

- 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:

$$Q(X_1, \ldots, X_k) = R_1(Vars_1) \wedge \ldots \wedge R_m(Vars_m)$$

- Let $w_1, w_2, \ldots, w_m$ be a fractional edge cover. Then the output size is no larger than:

$$|Q| \leq |R_1|^{w_1} \times \cdots \times |R_m|^{w_m}$$

Examples

Query	$w_1, w_2, \ldots, w_m$	$\lvert R_1\rvert^{w1} \times \ldots \times \lvert R_m\rvert^{wm}$	$\lvert Q\rvert \leq \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\lvert R\rvert \times \lvert S\rvert$	$\leq \lvert R\rvert \times \lvert S\rvert$

Examples

Query	$w_1, w_2, \ldots, w_m$	$\|R_1\|^{w1} \times \ldots \times \|R_m\|^{wm}$	$\|Q\| \leq \ldots$
R(x,y)∧S(y,z)	1,1	$\|R\| \times \|S\|$	$\leq \|R\| \times \|S\|$
R(x,y)∧S(y,z)∧T(z,x)			

Examples

Query	$w_1, w_2, \ldots, w_m$	$\lvert R_1\rvert^{w1} \times \ldots \times \lvert R_m\rvert^{wm}$	$\lvert Q\rvert \leq \ldots$
R(x,y)∧S(y,z)	1,1	$\lvert R\rvert \times \lvert S\rvert$	$\leq \lvert R\rvert \times \lvert S\rvert$
R(x,y)∧S(y,z)∧T(z,x)	½, ½, ½	$(\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}$	

Examples

Query	$w_1, w_2, \ldots, w_m$	$\lvert R_1\rvert^{w1} \times \ldots \times \lvert R_m\rvert^{wm}$	$\lvert Q\rvert \le \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\lvert R\rvert \times \lvert S\rvert$	$\le \lvert R\rvert \times \lvert S\rvert$
$R(x,y) \wedge S(y,z) \wedge T(z,x)$	½, ½, ½	$(\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}$	$\le \min((\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}, \lvert R\rvert \times \lvert S\rvert, \lvert R\rvert \times \lvert T\rvert, \lvert S\rvert \times \lvert T\rvert)$
	1,1,0	$\lvert R\rvert \times \lvert S\rvert$	
	1,0,1	$\lvert R\rvert \times \lvert T\rvert$	
	0,1,1	$\lvert S\rvert \times \lvert T\rvert$	

Examples

Query	$w_1, w_2, \ldots, w_m$	$\vert R_1\vert^{w1} \times \ldots \times \vert R_m\vert^{wm}$	$\vert Q\vert \le \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\vert R\vert \times \vert S\vert$	$\le \vert R\vert \times \vert S\vert$
$R(x,y) \wedge S(y,z) \wedge T(z,x)$	½, ½, ½	$(\vert R\vert \times \vert S\vert \times \vert T\vert)^{½}$	$\le \min((\vert R\vert \times \vert S\vert \times \vert T\vert)^{½}, \vert R\vert \times \vert S\vert, \vert R\vert \times \vert T\vert, \vert S\vert \times \vert T\vert)$
	1,1,0	$\vert R\vert \times \vert S\vert$	
	1,0,1	$\vert R\vert \times \vert T\vert$	
	0,1,1	$\vert S\vert \times \vert T\vert$	
$A(x,y,z) \wedge B(x,y,u) \wedge C(x,z,u) \wedge D(y,z,u)$			

Examples

Query	$w_1, w_2, \ldots, w_m$	$\|R_1\|^{w1} \times \ldots \times \|R_m\|^{wm}$	$\|Q\| \leq \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\|R\| \times \|S\|$	$\leq \|R\| \times \|S\|$
$R(x,y) \wedge S(y,z) \wedge T(z,x)$	½, ½, ½	$(\|R\| \times \|S\| \times \|T\|)^{½}$	$\leq \min((\|R\| \times \|S\| \times \|T\|)^{½}, \|R\| \times \|S\|, \|R\| \times \|T\|, \|S\| \times \|T\|)$
	1,1,0	$\|R\| \times \|S\|$	
	1,0,1	$\|R\| \times \|T\|$	
	0,1,1	$\|S\| \times \|T\|$	
$A(x,y,z) \wedge B(x,y,u) \wedge C(x,z,u) \wedge D(y,z,u)$	1/3, 1/3, 1/3, 1/3	$(\|A\| \times \|B\| \times \|C\| \times \|D\|)^{1/3}$	

Examples

Query	$w_1, w_2, \ldots, w_m$	$\|R_1\|^{w1} \times \ldots \times \|R_m\|^{wm}$	$\|Q\| \leq \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\|R\| \times \|S\|$	$\leq \|R\| \times \|S\|$
$R(x,y) \wedge S(y,z) \wedge T(z,x)$	½, ½, ½	$(\|R\| \times \|S\| \times \|T\|)^{½}$	$\leq \min((\|R\| \times \|S\| \times \|T\|)^{½}, \|R\| \times \|S\|, \|R\| \times \|T\|, \|S\| \times \|T\|)$
	1,1,0	$\|R\| \times \|S\|$	
	1,0,1	$\|R\| \times \|T\|$	
	0,1,1	$\|S\| \times \|T\|$	
$A(x,y,z) \wedge B(x,y,u) \wedge C(x,z,u) \wedge D(y,z,u)$	1/3, 1/3, 1/3, 1/3	$(\|A\| \times \|B\| \times \|C\| \times \|D\|)^{1/3}$	min(…)
	1,1,0,0	$\|A\| \times \|B\|$	
	1,0,1,0	$\|A\| \times \|C\|$	
	…	…	

Examples

Query	$w_1, w_2, \ldots, w_m$	$\|R_1\|^{w1} \times \ldots \times \|R_m\|^{wm}$	$\|Q\| \le \ldots$
R(x,y)∧S(y,z)	1,1	|R|×|S|	≤|R|×|S|
R(x,y)∧S(y,z)∧T(z,x)	½, ½, ½	(|R|×|S|×|T|)^{½}	≤min((|R|×|S|×|T|)^{½}, |R|×|S|, |R|×|T|, |S|×|T|)
	1,1,0	|R|×|S|	
	1,0,1	|R|×|T|	
	0,1,1	|S|×|T|	
A(x,y,z)∧B(x,y,u)∧C(x,z,u)∧ D(y,z,u)	1/3, 1/3, 1/3, 1/3	(|A|×|B|×|C|×|D|)$^{1/3}$	min(…)
	1,1,0,0	|A|×|B|	
	1,0,1,0	|A|×|C|	
	…	…	
R(x,y)∧S(y,z)∧T(z,u)∧K(u,v)			

Examples

Query	$w_1, w_2, \ldots, w_m$	$\lvert R_1\rvert^{w1} \times \ldots \times \lvert R_m\rvert^{wm}$	$\lvert Q\rvert \le \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\lvert R\rvert \times \lvert S\rvert$	$\le \lvert R\rvert \times \lvert S\rvert$
$R(x,y) \wedge S(y,z) \wedge T(z,x)$	½, ½, ½	$(\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}$	$\le \min((\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}, \lvert R\rvert \times \lvert S\rvert, \lvert R\rvert \times \lvert T\rvert, \lvert S\rvert \times \lvert T\rvert)$
	1,1,0	$\lvert R\rvert \times \lvert S\rvert$	
	1,0,1	$\lvert R\rvert \times \lvert T\rvert$	
	0,1,1	$\lvert S\rvert \times \lvert T\rvert$	
$A(x,y,z) \wedge B(x,y,u) \wedge C(x,z,u) \wedge D(y,z,u)$	1/3, 1/3, 1/3, 1/3	$(\lvert A\rvert \times \lvert B\rvert \times \lvert C\rvert \times \lvert D\rvert)^{1/3}$	min(…)
	1,1,0,0	$\lvert A\rvert \times \lvert B\rvert$	
	1,0,1,0	$\lvert A\rvert \times \lvert C\rvert$	
	…	…	
$R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,v)$	1,0,1,1	$\lvert R\rvert \times \lvert T\rvert \times \lvert K\rvert$	
	1,1,0,1	$\lvert R\rvert \times \lvert S\rvert \times \lvert K\rvert$	

Examples

<table>
<tr><th>Query</th><th>$w_1, w_2, \ldots, w_m$</th><th>$|R_1|^{w1} \times \ldots \times |R_m|^{wm}$</th><th>$|Q| \leq \ldots$</th></tr>
<tr><td>$R(x,y) \wedge S(y,z)$</td><td>1,1</td><td>$|R| \times |S|$</td><td>$\leq |R| \times |S|$</td></tr>
<tr><td rowspan="4">$R(x,y) \wedge S(y,z) \wedge T(z,x)$</td><td>½, ½, ½</td><td>$(|R| \times |S| \times |T|)^{1/2}$</td><td rowspan="4">$\leq \min((|R| \times |S| \times |T|)^{1/2},$ $|R| \times |S|, |R| \times |T|,$ $|S| \times |T|)$</td></tr>
<tr><td>1,1,0</td><td>$|R| \times |S|$</td></tr>
<tr><td>1,0,1</td><td>$|R| \times |T|$</td></tr>
<tr><td>0,1,1</td><td>$|S| \times |T|$</td></tr>
<tr><td rowspan="4">$A(x,y,z) \wedge B(x,y,u) \wedge C(x,z,u) \wedge D(y,z,u)$</td><td>1/3, 1/3, 1/3, 1/3</td><td>$(|A| \times |B| \times |C| \times |D|)^{1/3}$</td><td rowspan="4">min(…)</td></tr>
<tr><td>1,1,0,0</td><td>$|A| \times |B|$</td></tr>
<tr><td>1,0,1,0</td><td>$|A| \times |C|$</td></tr>
<tr><td>…</td><td>…</td></tr>
<tr><td rowspan="3">$R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,v)$</td><td>1,0,1,1</td><td>$|R| \times |T| \times |K|$</td><td rowspan="3"></td></tr>
<tr><td>1,1,0,1</td><td>$|R| \times |S| \times |K|$</td></tr>
<tr><td>1, ½, ½, 1</td><td>(no need; why?)</td></tr>
</table>

Examples

Query	$w_1, w_2, \ldots, w_m$	$\lvert R_1\rvert^{w1} \times \ldots \times \lvert R_m\rvert^{wm}$	$\lvert Q\rvert \le \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\lvert R\rvert \times \lvert S\rvert$	$\le \lvert R\rvert \times \lvert S\rvert$
$R(x,y) \wedge S(y,z) \wedge T(z,x)$	½, ½, ½	$(\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}$	$\le \min((\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}, \lvert R\rvert \times \lvert S\rvert, \lvert R\rvert \times \lvert T\rvert, \lvert S\rvert \times \lvert T\rvert)$
	1,1,0	$\lvert R\rvert \times \lvert S\rvert$	
	1,0,1	$\lvert R\rvert \times \lvert T\rvert$	
	0,1,1	$\lvert S\rvert \times \lvert T\rvert$	
$A(x,y,z) \wedge B(x,y,u) \wedge C(x,z,u) \wedge D(y,z,u)$	1/3, 1/3, 1/3, 1/3	$(\lvert A\rvert \times \lvert B\rvert \times \lvert C\rvert \times \lvert D\rvert)^{1/3}$	min(…)
	1,1,0,0	$\lvert A\rvert \times \lvert B\rvert$	
	1,0,1,0	$\lvert A\rvert \times \lvert C\rvert$	
	…	…	
$R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,v)$	1,0,1,1	$\lvert R\rvert \times \lvert T\rvert \times \lvert K\rvert$	$\min(\lvert R\rvert \times \lvert T\rvert \times \lvert K\rvert, \lvert R\rvert \times \lvert S\rvert \times \lvert K\rvert)$
	1,1,0,1	$\lvert R\rvert \times \lvert S\rvert \times \lvert K\rvert$	
	1, ½, ½, 1	(no need; why?)	

Upper Bound of a Query

Theorem $|Q| \leq \min_{w_1,\cdots,w_m} |R_1|^{w_1} \times \cdots \times |R_m|^{w_m}$

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

$\text{AGM}(Q) = \min_{w_1,\cdots,w_m} |R_1|^{w_1} \times \cdots \times |R_m|^{w_m}$

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.

$Q(x) = R(x) \wedge S(x)$

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

Generic Join – The Intersection

Intersection is the main building block of G.J.

$Q(x) = R(x) \wedge S(x)$

- 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.

$Q(x) = R(x) \wedge S(x)$

- 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}, … 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}, … are restricted to $x = v$

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

Generic Join Example

$Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$,
Assume $|R|=|S|=|T|=N$, then:

$|Q| \leq N^{3/2}$

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

Generic Join Example

$Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$,
Assume $|R|=|S|=|T|=N$, then:

$|Q| \leq N^{3/2}$

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

for a in A **do**

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

Generic Join Example

$Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$,
Assume $|R|=|S|=|T|=N$, then:

$|Q| \leq N^{3/2}$

$A = \Pi_x(R(x,y)) \cap \Pi_x(T(z,x))$
for a in A **do**
    /* compute $Q(a,y,z) = R(a,y) \wedge S(y,z) \wedge T(z,a)$ */
    $B = \Pi_y(R(a,y)) \cap \Pi_y(S(y,z))$
    for b in B **do**
    /* compute $Q(a,b,z) = R(a,b) \wedge S(b,z) \wedge T(z,a)$ */

Generic Join Example

$Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$,
Assume $|R|=|S|=|T|=N$, then:

$|Q| \leq N^{3/2}$

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

for a in A **do**

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

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

for b in B **do**

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

$C = \Pi_z(S(b,z)) \cap \Pi_z(T(z,a))$

for c in C **do**

output (a,b,c)

Generic Join Example

$Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$,
Assume $|R|=|S|=|T|=N$, then:

$|Q| \leq N^{3/2}$

$A = \Pi_x(R(x,y)) \cap \Pi_x(T(z,x))$
for a in A **do**
  /* compute $Q(a,y,z) = R(a,y) \wedge S(y,z) \wedge T(z,a)$ */
  $B = \Pi_y(R(a,y)) \cap \Pi_y(S(y,z))$
  for b in B **do**
  /* compute $Q(a,b,z) = R(a,b) \wedge S(b,z) \wedge T(z,a)$ */
    $C = \Pi_z(S(b,z)) \cap \Pi_z(T(z,a))$
    for c in C **do**
      output (a,b,c)

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

Discussion

- All relations need to be presorted, or indexed
- Runtime is guaranteed to be worst-case optimal, *no matter* what variable order we choose
- In practice, the variable order *does matter*; in class: discuss $R(x,y) \wedge S(y,z)$

Comparison to Naïve Nested Loop

Naïve nested loop:

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

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
      ...
```

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, … such that the size of the query's output is AGM(Q)

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

- In this case AGM(Q) = $N^{\min(w1+\ldots+wm)}$

Fractional Edge Covering Number

- The fractional edge covering number of a hypergraph is ρ^* = 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^{ρ^*}

Why?

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: $\sum_{x:\, x \in e} v_x \leq 1$

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: $\sum_x v_x \leq \sum_e w_e$

Fractional Vertex Packing

- A <u>*fractional vertex packing*</u> 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: $\sum_x v_x \leq \sum_e w_e$

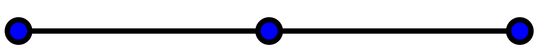
Theorem $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

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



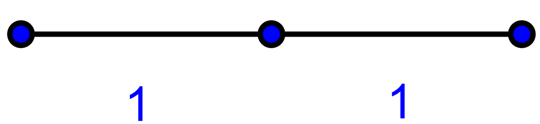
Fractional Vertex Packing

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

$\rho^* = 1$

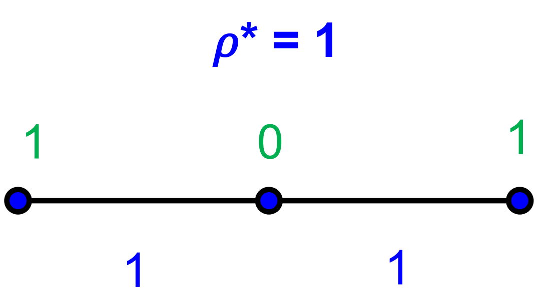


Fractional Vertex Packing

- A <u>*fractional vertex packing*</u> 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 $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

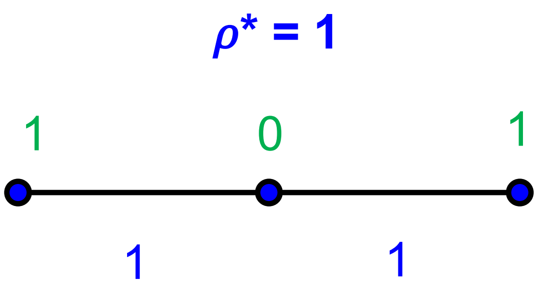


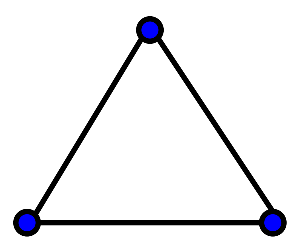
Fractional Vertex Packing

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





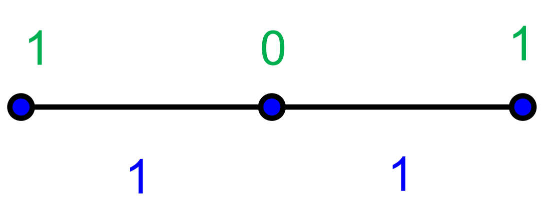
Fractional Vertex Packing

- 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: $\sum_{x:\, x \in e} v_x \leq 1$

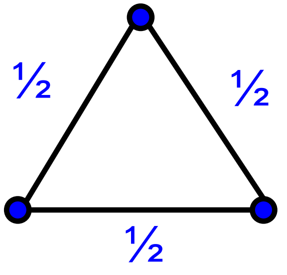
Fact For any v, w: $\sum_x v_x \leq \sum_e w_e$

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

$\rho^* = 1$



$\rho^* = 3/2$

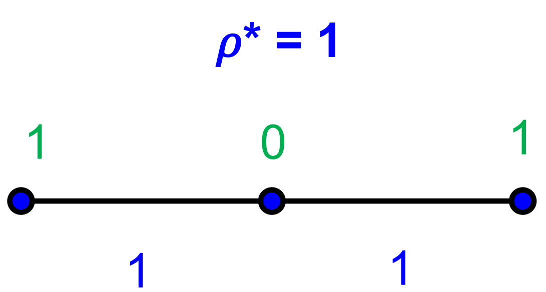


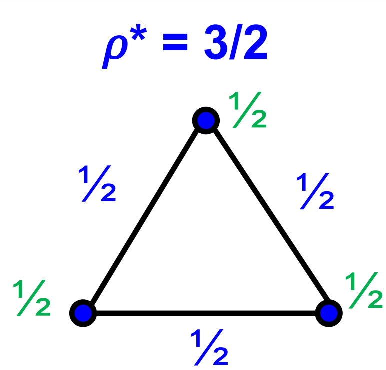
Fractional Vertex Packing

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



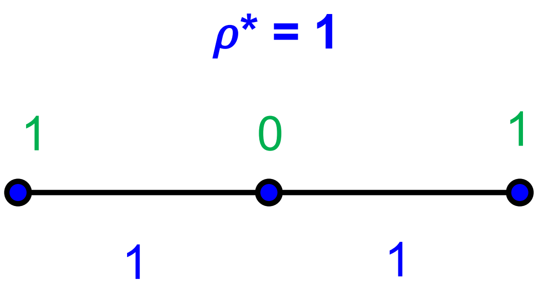


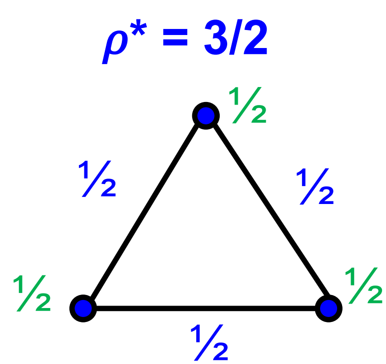
Fractional Vertex Packing

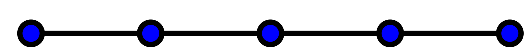
- A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers v_x, one for each node x, such that, for every edge e: $\sum_{x:\, x \in e} v_x \leq 1$

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

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







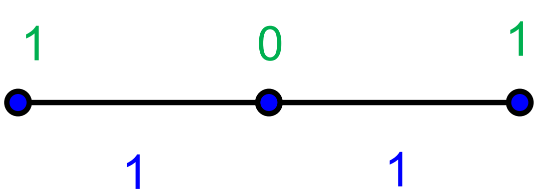
Fractional Vertex Packing

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

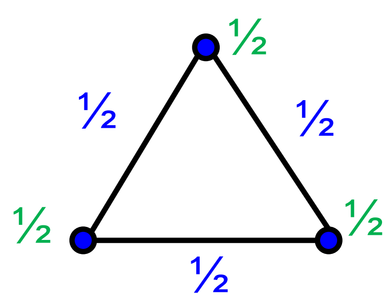
Fact For any v, w: $\sum_x v_x \leq \sum_e w_e$

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

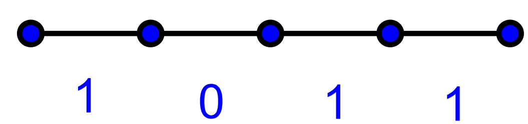
$\rho^* = 1$



$\rho^* = 3/2$



$\rho^* = 3$

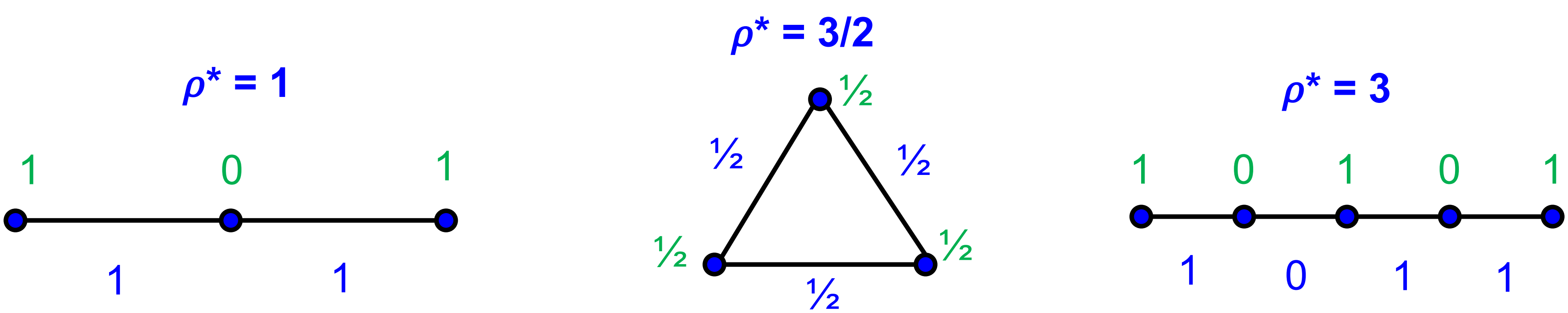


Fractional Vertex Packing

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



The Bound is Tight

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

The Bound is Tight

Fact Fix a fractional vertex packing v = $(v_x)_{x \in Nodes}$.
Then there exists a database such that
R$_1$| ≤ N, …, |R$_m$| ≤ N and $|Q| = N^{\Sigma_x v_x}$

Proof. For every relation R$_j$ with variables $x_{i_1}, x_{i_2}, \ldots$
define the instance $|R_j| = [N^{v_{i_1}}] \times [N^{v_{i_2}}] \times \cdots$
where [k] = {1,2,…,k}.

The Bound is Tight

Fact Fix a fractional vertex packing v = $(v_x)_{x \in Nodes}$.
Then there exists a database such that
$R_1| \le N$, …, $|R_m| \le N$ and $|Q| = N^{\Sigma_x v_x}$

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

The Bound is Tight

Fact Fix a fractional vertex packing v = $(v_x)_{x \in Nodes}$.
Then there exists a database such that
$R_1| \leq N$, …, $|R_m| \leq N$ 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 $|R_j| = [N^{v_{i_1}}] \times [N^{v_{i_2}}] \times \cdots$
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

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

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

Example

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

- Assume $|R|=|S|=|T|=N$.
- Optimal vertex packing: $v_x = v_y = v_z = 1/2$

Example

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

- Assume $|R|=|S|=|T|=N$.
- Optimal vertex packing: $v_x = v_y = v_z = 1/2$
- Define: $D_x = [N^{½}] = \{1, 2, \ldots, N^{½}\}$
  $D_y = [N^{½}]$
  $D_z = [N^{½}]$

Example

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

- Assume $|R|=|S|=|T|=N$.
- Optimal vertex packing: $v_x = v_y = v_z = 1/2$
- Define: $D_x = [N^{½}] = \{1, 2, \ldots, N^{½}\}$
  $D_y = [N^{½}]$
  $D_z = [N^{½}]$
- Define $R = D_x \times D_y$, $S = D_y \times D_z$, $T = D_z \times D_x$.

Example

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

- 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

$R(X,Y) \wedge S(Y,Z)$, $|R|, |S| \leq N$

- 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 AGM($Q^{expanded}$)

Examples

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z)$

Examples

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z)$

- $Q^{exp}(X,Y,Z) = R(X,Y,Z) \wedge S(Y,Z)$,

Examples

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z)$

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

Examples

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z)$

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

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z) \wedge T(Z,X)$

Examples

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z)$

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

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z) \wedge T(Z,X)$

- $Q^{exp}(X,Y,Z) = R(X,Y,Z) \wedge S(Y,Z) \wedge T(Z,X)$

Examples

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z)$

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

$Q(X,Y,Z) = R(X,Y) \wedge S(\underline{Y},Z) \wedge T(Z,X)$

- $Q^{exp}(X,Y,Z) = R(X,Y,Z) \wedge S(Y,Z) \wedge T(Z,X)$
- Edge covers: 1,0,0 or 0,1,1
- AGM(Q^{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